



UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

**Classification Techniques
for
Time Series and Functional Data**

Autor:

David Casado de Lucas

Directores:

Andrés M. Alonso y Juan Romo

Departamento de Estadística

Leganés, julio del 2010

TESIS DOCTORAL

Classification Techniques for Time Series and Functional Data

Autor: David Casado de Lucas

Directores: Andrés M. Alonso y Juan Romo

Firma del Tribunal Calificador:

Presidente: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Vocal: (Nombre y apellidos)

Secretario: (Nombre y apellidos)

Firma

Calificación:

Leganés, de de

To my family

Acknowledgements

I want to thank my family for the absolute confidence they have always placed in me (even when they did not understand why my —our— salary and holidays were not “those teachers usually have”). I would also like to thank my advisors, Andrés and Juan, and the Universidad Carlos III de Madrid.

Preface

Here is the research work I have done, in collaboration with my advisors, over several academic years at the Universidad Carlos III de Madrid. I would like to say I am very pleased to have had the opportunity of helping with the teaching tasks at the Departamento de Estadística.

This text is organized as follows. Chapter one starts with the basic mathematical notation. Next sections include the main definitions and results that appear in the rest of the thesis. Finally, general ideas on the statistical problem of *classification* are provided, just before the exposition of our proposals. Chapters two and three are devoted, respectively, to the time series and the functional data classification methods. These chapters can be independently consulted. Several possible extensions and forthcoming work are proposed in chapter four. Finally, some conclusions are summarised.

This structure allows independent access to the contents, and especially to chapters two and three (our contribution, apart from the redaction of the text). In both chapter one and the appendices, the theory included is the “minimum necessary” while being, at the same time, self-contained.

Contents

Preface	ii
Resumen	1
Abstract	3
1 Introduction	5
1.1 Notation	5
1.2 Stochastic Vectors	6
1.2.1 Models	6
1.2.2 Multivariate Data	7
1.2.3 Statistical Inference	8
1.2.4 Addendum: Compound Variable Geometry	9
1.3 Stochastic Processes	11
1.3.1 Models	11
1.3.2 Time Series Data	16
1.3.3 Statistical Inference	16
1.3.4 Addendum: Locally Stationary Processes	18
1.4 Stochastic Functions	21
1.4.1 Models	21
1.4.2 Functional Data	23
1.4.3 Statistical Inference	25
1.4.4 Addendum: Functional Depth	26
1.5 The Classification Problem	28
1.5.1 Theoretical Rule	29

1.5.2	Approximate Rule	30
1.5.3	Sample Rule	31
1.5.4	Stochastic Rule	32
1.5.5	Asymptotic Rule	32
1.5.6	Example	34
1.5.7	Our Classification Frameworks	35
2	Time Series Classification	37
2.1	Introduction	37
2.2	The Classification Method	39
2.2.1	The Integrated Periodogram	39
2.2.2	Classifying Functions	40
2.2.3	The Algorithm	41
2.3	Robust Version	42
2.3.1	New Reference Function	42
2.3.2	New Algorithm	42
2.4	Simulation Results	43
2.5	Real Data Example	58
2.5.1	Explosions and Earthquakes Data	58
2.6	Conclusions	59
3	Functional Data Classification	61
3.1	Introduction	61
3.2	The Classification Method	64
3.2.1	The Optimization Problem	64
3.2.2	The Discriminant Function	65
3.2.3	The Classification	69
3.2.4	Our Discriminant Variables	70
3.2.5	The Algorithm	71
3.2.6	Weighted Semidistances or Distances	72
3.3	Simulation Results	73
3.4	Real Data Examples	78

3.4.1	Spectrometric Data	79
3.4.2	Growth Data	82
3.5	Conclusions	84
4	Extensions and Further Work	85
4.1	Time Series Method	85
4.1.1	More than Two Populations	85
4.1.2	Clustering	86
4.1.3	Other Depth Definitions	86
4.2	Functional Data Method	86
4.2.1	Classical Assumptions	86
4.2.2	Additional Constraint Embedding	87
4.2.3	Additional Constraint Avoidance	88
4.2.4	Transformation Importance	89
4.2.5	Distance Importance	90
4.2.6	Several Discriminant Functions	90
4.2.7	Other Classification Methods	91
	Conclusions	92
	A Vector Analysis	94
	B Fisher's Discriminant Analysis	99
B.1	Motivation	99
B.1.1	Parameter Estimation	100
B.1.2	Variability Information	100
B.1.3	Splitting Criterion	103
B.1.4	Case $q = 1$: One Function	103
B.2	The Optimization Problem	104
B.2.1	Equivalent Problems	105
B.2.2	Case $K = 2$: Two Populations	107
B.2.3	Interpretation of the Coefficients	107
B.3	The Discriminant Function	109

B.4	The Classification	109
C	Optimization Theory	112
	Karush-Kuhn-Tucker Conditions	123
	References	125

Resumen

Esta tesis doctoral introduce técnicas de clasificación para datos dependientes y funcionales. Se proponen sendos métodos para clasificar series temporales y datos funcionales. Aunque este trabajo involucra varios tipos de datos, los datos funcionales desempeñan un papel central. Un aspecto importante de ambas metodologías de clasificación es que los problemas originales no se afrontan directamente: el problema de series temporales se reescribe como un problema de datos funcionales, mientras que el problema de datos funcionales se resuelve utilizando una técnica multivariante. Por otro lado, también es interesante advertir el papel diferente que los datos funcionales tienen en las dos propuestas que se presentan: en el problema de series temporales se construyen estimadores funcionales, mientras que en el problema de datos funcionales las curvas son los datos primarios.

Para el problema de clasificación de series temporales, se consideran sus periodogramas integrados en lugar de las propias series. Un nuevo elemento se asigna al grupo que hace mínima la distancia desde su periodograma integrado hasta la media de los periodogramas integrados del grupo. Aunque el periodograma está definido sólo para series temporales estacionarias, es todavía posible aplicar la metodología a series no estacionarias considerando estos periodogramas localmente. Por último, se aplica la profundidad de datos funcionales para hacer la clasificación robusta.

Por un lado, la clasificación de datos funcionales surge de manera natural en el marco de trabajo anterior. Por otro lado, se sugiere el problema de seleccionar la forma más apropiada en la que expresar los datos: las funciones originales, sus integrales o sus derivadas. Sin pérdida de generalidad, este segundo problema se formula equivalentemente en términos de las funciones y de sus derivadas de distintos órdenes, sin integrales. En esta tesis se propone un tipo de metodología para hacer frente a estos dos problemas a la vez. Siguiendo de nuevo el criterio de clasificar una curva utilizando las distancias desde la función o sus derivadas hasta funciones

representativas de cada grupo (normalmente la media), nuestro método propone la combinación de estas distancias. La propuesta trabaja con una variable multivariante definida en términos de las distancias. Además, se tiene así una forma automática de ordenar por poder discriminante las funciones originales y sus derivadas.

Palabras clave: clasificación, datos de series temporales, periodograma integrado, datos funcionales, profundidad, datos multivariantes, análisis discriminante, distancias ponderadas.

Abstract

The main subject of this doctoral thesis is to develop classification techniques for dependent and functional data. Methods for classifying time series and functional data are proposed. Although this work involves several type of data, the functional data play a central role. An important point of both classification methodologies is that the original problems are not directly dealt with: the time series problem is rewritten as a functional data problem while the functional data problem is solved using a multivariate technique. It is worthwhile noticing, however, the different role of the functional data in the two forthcoming proposals: in the time series problem functional estimators are constructed, while in the functional data problem curves are the primary data.

For the classification of time series, their integrated periodograms are considered instead. After this, a new element is assigned to the group minimizing the distance from its integrated periodogram to the group mean of integrated periodograms. Although the periodogram is defined only for stationary time series, the application of the methodology to nonstationary series is still possible by computing these periodograms locally. Finally, functional data depth is applied to make the classification robust.

On the one hand, the classification of functional data arises naturally in the previous framework. On the other hand, the problem of selecting the more appropriate form to express the data is suggested: crude functions, their integrals or their derivatives. Without loss of generality, this second problem is equivalently formulated in terms of functions and their derivatives of different order, without integrals. In this thesis, a single methodology is proposed to cope with these two tasks at the same time. Following the same criterion of classifying a curve by using the distances from the function or its derivatives to group representative (usually the mean) functions or their derivatives, the combination of those distances is proposed in our method. The proposal works with a multivariate variable defined in terms of the distances. Besides, an automatic form of ranking the original functions and their derivatives by discriminant power is obtained.

Key words: classification, time series data, integrated periodogram, functional data, depth, multivariate data, discriminant analysis, weighted distances.

Chapter 1

Introduction

Summary: The different types of data involved in this thesis are explained with the minimum necessary extension. Next sections start with the theoretical models frequently used to fit the data. Knowing these models is essential to understand the statistical methods proposed later. Moreover, the structure “models \rightarrow data \rightarrow inference” allows for the subsections on inference to be after the subsections on the data. Finally, the classification problem is presented in a general form.

Key words: multivariate data, time series, functional data, supervised classification.

1.1 Notation

The following mathematical notation will be used throughout this document:

X and x for a univariate variable

\mathbf{X} and \mathbf{x} for a multivariate variable

X_t and x_t for a (discrete-time) sequence of variables

$X(t)$ and $x(t)$ for a (continuous-time) family of variables

$\mathcal{X}(t)$ and $\chi(t)$ for a function (depending on the variable t)

The letter Y is used for mathematical objects of the same kind that —but with some dependence on— the corresponding objects denoted with X .

In addition, for samples, the subindex e denotes the e -th element of a sample of size n , and, when there are K different populations or groups, the superindex $^{(k)}$ denotes the k -th population or group.

For a general matrix \mathbf{M} of size $n_1 \times n_2$, the following notation will be used sometimes:

$$\mathbf{M} = \begin{pmatrix} m_{11} & \cdots & m_{1n_2} \\ \vdots & \ddots & \vdots \\ m_{n_1 1} & \cdots & m_{n_1 n_2} \end{pmatrix} = (m_{ij})_{i,j},$$

where $i = 1, \dots, n_1$, $j = 1, \dots, n_2$. Finally,

$D^i \mathcal{X}(t)$ and $D^i \chi(t)$ denote the i -th derivative,

with respect to the real variable t , where $D^0 \mathcal{X}(t) \equiv \mathcal{X}(t)$ and $D^0 \chi(t) \equiv \chi(t)$.

1.2 Stochastic Vectors

1.2.1 Models

Definition 1 A multivariate random variable or random vector of dimension p is defined as

$$\mathbf{X} = (X_1, \dots, X_p)^t, \quad (1.1)$$

where X_i are simultaneous univariate random variables (defined in the same probability space).

The previous vector can be represented through the following application

$$\begin{aligned} \mathbf{X} : \quad \Omega &\longrightarrow \mathbb{R}^p \\ \omega &\longrightarrow \mathbf{X}(\omega) \end{aligned}$$

That is, for fixed ω a value in \mathbb{R}^p is obtained, while a random variable is obtained when the subindex—in the vector—is fixed.

Definition 2 The mean vector of the multivariate random variable \mathbf{X} is defined as the (column, in this case) vector

$$\mu_{\mathbf{X}} = (\mu_1, \dots, \mu_p)^t = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_p))^t = \mathbb{E}(\mathbf{X}). \quad (1.2)$$

Definition 3 The covariance matrix of the variable \mathbf{X} is defined as

$$\begin{aligned} \Sigma_{\mathbf{X}} &= (\sigma_{ij})_{i,j} = (\mathbb{E}((X_i - \mu_i)(X_j - \mu_j)))_{i,j} \\ &= \mathbb{E}(((X_i - \mu_i)(X_j - \mu_j))_{i,j}) = \mathbb{E}((\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^t). \end{aligned} \quad (1.3)$$

Definition 4 A univariate compound random variable is defined as

$$Y = a_1X_1 + \dots + a_pX_p = \mathbf{a}^t\mathbf{X}, \quad (1.4)$$

where $\mathbf{a} = (a_1, \dots, a_p)^t \in \mathbb{R}^p$.

Definition 5 A multivariate compound random variable of dimension q is defined as

$$\mathbf{Y} = (Y_1, \dots, Y_q)^t, \quad (1.5)$$

with

$$Y_j = a_{1j}X_1 + \dots + a_{pj}X_p, \quad j = 1, \dots, q \quad (1.6)$$

or, in matrix notation,

$$\mathbf{Y} = \mathbf{A}^t\mathbf{X}, \quad (1.7)$$

where $\mathbf{A} = (a_{ij})_{i,j}$ is the $p \times q$ matrix of the coefficients.

For the previous definitions, it holds that

Proposition 1

$$\mu_{\mathbf{Y}} = \mathbf{A}^t\mu_{\mathbf{X}}, \quad (1.8)$$

and, for the particular univariate case, $\mu_Y = \mathbf{a}^t\mu_{\mathbf{X}}$.

Proposition 2

$$\Sigma_{\mathbf{Y}} = \mathbf{A}^t\Sigma_{\mathbf{X}}\mathbf{A}, \quad (1.9)$$

and, for the particular univariate case, $\Sigma_Y = \mathbf{a}^t\Sigma_{\mathbf{X}}\mathbf{a}$.

1.2.2 Multivariate Data

Definition 6 A multivariate variable of dimension p is defined as the vector

$$\mathbf{x} = (x_1, \dots, x_p)^t, \quad (1.10)$$

where x_i are simultaneous univariate numeric variables.

In this thesis, numeric variables are quantitative and continuous, and they can be thought of as realizations of random variables. This definition makes it clear that the termed *Multivariate Analysis* deals with the simultaneous relationships among the variables of the vector. The compound variables are defined as

Definition 7 A univariate compound variable is defined as

$$y = a_1x_1 + \dots + a_px_p = \mathbf{a}^t\mathbf{x}, \quad (1.11)$$

where $\mathbf{a} = (a_1, \dots, a_p)^t \in \mathbb{R}^p$.

Definition 8 A multivariate compound variable of dimension q is defined as

$$\mathbf{y} = (y_1, \dots, y_q)^t, \quad (1.12)$$

with

$$y_j = a_{1j}x_1 + \dots + a_{pj}x_p, \quad j = 1, \dots, q \quad (1.13)$$

or, in matrix notation,

$$\mathbf{y} = \mathbf{A}^t\mathbf{x}, \quad (1.14)$$

where $\mathbf{A} = (a_{ij})_{i,j}$ is the $p \times q$ matrix of the coefficients.

1.2.3 Statistical Inference

A sample of size n of multivariate data can be expressed in a matrix:

$$(\mathbf{x}_1, \dots, \mathbf{x}_n) = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{pmatrix} = (x_{ij})_{i,j}, \quad (1.15)$$

where \mathbf{x}_e , the e -th column of the matrix, contains the e -th element of the sample of size n . It is supposed that the multivariate variables \mathbf{x}_e are also simultaneous, that is, in this situation there is no interest in the evolution through time. Thus, some well-known results on the parameter estimation of the models from the samples of data are the following (see, for example, section 2.8 of Mardia et al. [1979]).

MEAN VECTOR

- The quantity $\mathbb{E}(X_i)$ is estimated by $\bar{x}_i = \frac{1}{n} \sum_{e=1}^n x_{ie}$.
- Then, $\mu_{\mathbf{x}}$ is estimated by $\bar{\mathbf{x}} = \frac{1}{n} \sum_{e=1}^n \mathbf{x}_e$.
- For the multivariate compound variable \mathbf{Y} , the mean $\mu_{\mathbf{Y}}$ is estimated by $\bar{\mathbf{y}} = \mathbf{A}^t\bar{\mathbf{x}}$. For the univariate compound variable Y , the mean μ_Y is estimated by $\bar{y} = \mathbf{a}^t\bar{\mathbf{x}}$.

COVARIANCE MATRIX

- The matrix $\Sigma_{\mathbf{x}}$ is estimated with bias by $\hat{\Sigma}_{\mathbf{x}} = \frac{1}{n} \sum_{e=1}^n (\mathbf{x}_e - \bar{\mathbf{x}})(\mathbf{x}_e - \bar{\mathbf{x}})^t$ (see expression [3.20] for an equivalent formula of this estimator).
- The matrix $\Sigma_{\mathbf{x}}$ is estimated without bias by $\mathbf{S}_{\mathbf{x}} = \frac{n}{n-1} \hat{\Sigma}_{\mathbf{x}}$.
- For the variable \mathbf{Y} , the matrix $\Sigma_{\mathbf{Y}}$ is estimated with bias by $\hat{\Sigma}_{\mathbf{Y}} = \mathbf{A}^t \hat{\Sigma}_{\mathbf{x}} \mathbf{A}$. For the variable Y , the matrix Σ_Y is estimated with bias by $\hat{\Sigma}_y = \mathbf{a}^t \hat{\Sigma}_{\mathbf{x}} \mathbf{a}$.
- For the variable \mathbf{Y} , the matrix $\Sigma_{\mathbf{Y}}$ is estimated without bias by $\mathbf{S}_{\mathbf{Y}} = \mathbf{A}^t \mathbf{S}_{\mathbf{x}} \mathbf{A}$. For the variable Y , the matrix Σ_Y is estimated without bias by $S_y = \mathbf{a}^t \mathbf{S}_{\mathbf{x}} \mathbf{a}$.

1.2.4 Addendum: Compound Variable Geometry

In this subsection, the geometric interpretation of the univariate compound variable is highlighted.

$$y = a_1 x_1 + \dots + a_p x_p = \mathbf{a}^t \mathbf{x}, \quad (1.16)$$

where $\mathbf{a} = (a_1, \dots, a_p)^t \in \mathbb{R}^p$ is the parameter.

GEOMETRY IN \mathbb{R}^p

Let \mathbf{x} and \mathbf{a} be (column) vectors, both with origin in $\mathbf{0}$; then, the projection of \mathbf{x} onto the line —one-dimensional linear subspace— determined by \mathbf{a} is obtained as follows. Let $V_{\mathbf{a}}$ be the linear subspace of \mathbb{R}^p generated by \mathbf{a} ; then any $\mathbf{x} \in \mathbb{R}^p$ can be written uniquely as the sum of the projections on $V_{\mathbf{a}}$ and its complementary $V_{\mathbf{a}}^{\perp}$, that is, symbolically, $\mathbb{R}^p = V_{\mathbf{a}} \oplus V_{\mathbf{a}}^{\perp}$. Since

$$\mathbf{x} = Proj_{V_{\mathbf{a}}}(\mathbf{x}) + Proj_{V_{\mathbf{a}}^{\perp}}(\mathbf{x}) = c\mathbf{a} + Proj_{V_{\mathbf{a}}^{\perp}}(\mathbf{x}), \quad (1.17)$$

after premultiplying by \mathbf{a} ,

$$\langle \mathbf{a}, \mathbf{x} \rangle = c \langle \mathbf{a}, \mathbf{a} \rangle, \quad (1.18)$$

so, as $\langle \mathbf{a}, \mathbf{a} \rangle \neq 0$,

$$Proj_{V_{\mathbf{a}}}(\mathbf{x}) = c\mathbf{a} = \frac{\langle \mathbf{a}, \mathbf{x} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle} \mathbf{a} = \left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_e}, \mathbf{x} \right\rangle \frac{\mathbf{a}}{\|\mathbf{a}\|_e} = \left(\frac{\mathbf{a}^t}{\|\mathbf{a}\|_e} \mathbf{x} \right) \frac{\mathbf{a}}{\|\mathbf{a}\|_e}, \quad (1.19)$$

where $\|\cdot\|_e$ denotes the Euclidean norm in \mathbb{R}^p and $\langle \cdot, \cdot \rangle$ denotes the inner product. Then, the parameter \mathbf{a} of the compound variable can be interpreted geometrically as a parameter determining the direction into which \mathbf{x} is projected.

The linear subspace and the vectors \mathbf{x} and $Proj_{V_{\mathbf{a}}}(\mathbf{x})$ can be “represented” (literally only for low dimensions) in \mathbb{R}^p .

Alternatively, if there is no interest in the interpretation of the line as a subspace of \mathbb{R}^p , this line can be seen as an independent axis.

GEOMETRY IN \mathbb{R}

For compound variables, the interest relies usually on the module of the projection, that is, in the part

$$\left\langle \frac{\mathbf{a}}{\|\mathbf{a}\|_e}, \mathbf{x} \right\rangle = \frac{\mathbf{a}^t}{\|\mathbf{a}\|_e} \mathbf{x}. \quad (1.20)$$

In addition, if there is no interest in the scale factor of the projection either, a characterization is provided by the numeric factor

$$\mathbf{a}^t \mathbf{x} = y. \quad (1.21)$$

Remark 1 Note that this approach allows the use of the expression “the premultiplication \mathbf{a}^t projects \mathbf{x} into a one-dimensional space”.

Another possible geometric interpretation of the compound variable can be obtained by thinking in a higher-dimensional linear space.

GEOMETRY IN \mathbb{R}^{p+1}

The equation (1.21) can be written as

$$\mathbf{a}^t \mathbf{x} - y = 0, \quad (1.22)$$

and, thinking now in the $(p+1)$ -dimensional linear space determined by the variables (\mathbf{x}, y) , it is clear that the hyperplane (1.22), with director vector $(\mathbf{a}^t, -1)$, will be the way of projecting the points of a multivariate sample \mathbf{x} into the last axis, the y axis.

These geometrical interpretations will be reviewed several times in chapter 3 and appendix B, since the choice of \mathbf{a} , from the information provided by multivariate data, will be the main objective there. Selecting \mathbf{a} for the compound variable can be interpreted geometrically as selecting —the direction of— a one-dimensional subspace into which the multivariate data will be projected. This choice is usually guided by some explicit criterion. Although in the theoretical compound variable \mathbf{x} is a multivariate variable and \mathbf{a} is a multivariate parameter, in the empirical selection of the direction of the one-dimensional linear subspace, \mathbf{a} will be temporally considered a multivariate variable. Finally, let us note that making any restriction on the value of \mathbf{a} is equivalent to making restrictions on the direction into which the compound variable project the data.

1.3 Stochastic Processes

In this section, only discrete-time univariate stochastic processes are considered.

1.3.1 Models

Definition 9 *A discrete-time stochastic process is a sequence of random variables (defined in the same probability space and taking values in the same state space S),*

$$(X_t, t \in \mathbb{Z}), \tag{1.23}$$

where the index t represents time.

In the previous process, for fixed t the quantity X_t is a random variable, while for fixed ω a *trajectory* is obtained, as represented through the following application

$$\begin{array}{ccc} X_t : & \Omega & \longrightarrow \mathbb{R}^\infty \\ & \omega & \longrightarrow X_t(\omega) \end{array} .$$

Note that now the variables are not simultaneous, as those of the multivariate vector (1.1). This mathematical object is defined to study dynamic (univariate) changes in time. On the other hand, another possibility could be to consider multivariate stochastic processes $(\mathbf{X}_t, t \in \mathbb{Z})$, where each \mathbf{X}_t is a vector of simultaneous variables.

TWO DOMAINS

Stochastic processes can be studied from the *time domain* or, alternatively or additionally, from the *spectral domain*. Time domain uses position or time as index, and the autocovariance — or, equivalently, autocorrelation— function is the natural tool for studying the evolution in the time domain. The previous definition corresponded to the time domain, but the *Fourier Analysis* allows the use of the frequency as variable. The *Spectral Analysis* is the adaptation of the Fourier analysis to deal with stochastic —rather than deterministic— functions of time.

In order to handle processes, some additional structure is assumed under the name of *stationarity*. It implies homogeneity in the time domain, in the form of autocovariance function invariant under time shifts. This similarity with the periodicity of functions will allow the decomposition of processes, under some conditions, in terms of regular underlying oscillations whose magnitudes are random variables; that is, a decomposition into the sum of uncorrelated periodic components. The *spectrum*, the set of frequencies of oscillations, is the natural mathematical tool in the frequency domain.

Definition 10 *The stochastic process (X_t) is strongly stationary if the random vectors,*

$$(X_{t_1}, \dots, X_{t_d}) \quad \text{and} \quad (X_{t_1+s}, \dots, X_{t_d+s}) \quad (1.24)$$

have the same joint distribution for all t_1, \dots, t_d and for all $s > 0$.

Thus, the finite-dimensional distributions of an strongly stationary process are invariant under time shifts. These distributions characterize the whole process. For processes such that $\text{Var}(X_t) < \infty, \forall t \in \mathbb{Z}$, a weaker condition —implied by the previous— is presented in the following definition.

Definition 11 *The stochastic process (X_t) is weakly (also termed second-order or covariance) stationary if $\mathbb{E}(|X_t|^2) < \infty$, for all $t \in \mathbb{Z}$, and*

$$1. \mathbb{E}(X_{t_1}) = \mathbb{E}(X_{t_2}),$$

$$2. \text{cov}(X_{t_1}, X_{t_2}) = \text{cov}(X_{t_1+s}, X_{t_2+s}),$$

for all $t_1, t_2, s \in \mathbb{Z}$ with $s > 0$.

Once the stationarity is imposed, the following definitions make sense.

Definition 12 The mean function of a weakly stationary process (X_t) is defined as

$$\mu_t = \mathbb{E}(X_t), \quad t \in \mathbb{Z}. \quad (1.25)$$

Definition 13 The autocovariance function of a weakly stationary process (X_t) is defined as

$$\sigma_s = \text{cov}(X_t, X_{t+|s|}), \quad s \in \mathbb{Z}, \quad (1.26)$$

where the *covariance* of any two variables (with finite mean) is defined as

$$\text{cov}(X_1, X_2) = \mathbb{E}([X_1 - \mathbb{E}(X_1)][X_2 - \mathbb{E}(X_2)]). \quad (1.27)$$

The *variance* of any variable X is defined as $\text{var}(X) = \text{cov}(X, X)$.

From these definitions, it follows that the process (X_t) is weakly stationary if and only if

1. It has constant mean.
2. Its autocovariance function is invariant under time shifts.

Stationary processes can be described in terms of the autocovariance function (it contains enough information) or, equivalently, in terms of the following rescaled function.

Definition 14 The autocorrelation function of a weakly stationary process (X_t) is defined as

$$\rho_s = \frac{\text{cov}(X_0, X_{|s|})}{\sqrt{\text{var}(X_0)\text{var}(X_{|s|})}} = \frac{\sigma_{|s|}}{\sigma_0}, \quad s \in \mathbb{Z} \quad (1.28)$$

whenever $\sigma_0 = \text{var}(X_{|s|}) > 0$.

Classical results linking both domains are provided by the Fourier analysis.

Theorem 1 - Wold's Theorem. A necessary and sufficient condition for a sequence $\{\rho_s\}$ to be the autocorrelation function for some discrete time stationary process (X_t) is that there exists a function $F(\lambda)$, having the properties of a distribution function on the interval $(-\pi, +\pi)$, (i.e. $F(-\pi) = 0$, $F(+\pi) = 1$ and $F(\lambda)$ is nondecreasing), such that

$$\rho_s = \int_{-\pi}^{+\pi} e^{is\lambda} dF(\lambda), \quad s \in \mathbb{Z}. \quad (1.29)$$

Proof. See section 4.8.3 of Priestley (1981). □

Definition 15 If the autocorrelation ρ_s satisfies (1.29) then $F(\lambda)$ is called the spectral distribution function of the process.

Theorem 2 Any integrated spectrum $F(\lambda)$ can be written in the form,

$$F(\lambda) = c_1 F_1(\lambda) + c_2 F_2(\lambda) + c_3 F_3(\lambda), \quad (1.30)$$

where

1. $c_i \geq 0$, $i = 1, 2, 3$ and $c_1 + c_2 + c_3 = 1$
2. $F_i(\lambda) \geq 0$, $i = 1, 2, 3$ are distribution functions of the following types;
 - (a) $F_1(\lambda)$ is absolutely continuous with derivative $F_1'(\lambda)$ which exists for almost all λ , and the density function $f_1(\lambda)$, which is such that $F_1(\lambda) = \int_{-\pi}^{\lambda} f_1(h)dh$, exists for all λ .
 - (b) $F_2(\lambda)$ is a step function with steps $\{p_s\}$ at points λ_s , say, $s = 1, 2, \dots$, and $\sum_s p_s = 1$.
 - (c) $F_3(\lambda)$ is a “singular” function with zero derivative almost everywhere.

Proof. See section 4.9 of Priestley (1981). □

Remark 2 The singular part $F_3(\lambda)$ in (1.30) is highly pathological and usually ignored.

Definition 16 A stochastic process (X_t) is said to have purely continuous spectrum if $F \equiv F_1$ in expression (1.30), that is, the other two parts are null.

Theorem 3 - Spectral Theorem. If (X_t) is a discrete-time stationary process with zero mean, unit variance, and spectral distribution function $F(\lambda)$, there exists a complex-valued process $\mathbf{S} = (S(\lambda), -\pi < \lambda \leq +\pi)$ such that

$$X_t = \int_{-\pi}^{+\pi} e^{it\lambda} dS(\lambda), \quad t \in \mathbb{Z}. \quad (1.31)$$

Furthermore, \mathbf{S} has orthogonal increments and

$$\mathbb{E}(|S(v) - S(u)|^2) = F(v) - F(u) \quad \text{for } u \leq v. \quad (1.32)$$

Proof. See section 9.4 of Grimmett and Stirzaker (2001). □

PURELY CONTINUOUS SPECTRUM

A purely indeterministic discrete-time stationary process (X_t) verifies that $\sum_{s=1}^{+\infty} |\rho_s| < \infty$ and, as a consequence, has purely continuous spectrum. In this case, the spectral distribution function $F(\lambda)$ is absolutely continuous and has density function $f(\lambda)$, named as follows:

Definition 17 The function $f(\lambda)$ is termed spectral density function.

It holds that

$$F(\lambda) = \int_{-\pi}^{\lambda} f(h)dh, \quad \lambda \in [-\pi, +\pi], \quad (1.33)$$

so

$$F'(\lambda) = \frac{d}{d\lambda}F(\lambda) = f(\lambda), \quad \lambda \in [-\pi, +\pi]. \quad (1.34)$$

Some other consequences are that the expression (1.29) becomes

$$\rho_s = \int_{-\pi}^{+\pi} e^{is\lambda} f(\lambda) d\lambda, \quad s \in \mathbb{Z}, \quad (1.35)$$

and that the following theorem holds.

Theorem 4 Let $\{\rho_s\}$ be the autocorrelation function (sequence) of a stationary sequence. If the function $F(\lambda)$ in (1.29) is differentiable with derivative $f(\lambda)$, then

$$f(\lambda) = \frac{1}{2\pi} \sum_{s=-\infty}^{+\infty} \rho_s e^{-is\lambda} \quad (1.36)$$

at any point $\lambda \in [-\pi, +\pi]$ where $f(\lambda)$ is differentiable.

Proof.

See section 9.3 of Grimmett and Stirzaker (2001). □

Remark 3 It is important not to confuse the expression “purely indeterministic”, that is, without deterministic part at all, with the expression “purely random”, which is frequently used for referring to the *white noise* or to an independent or uncorrelated sequence or family of variables.

Remark 4 The decomposition of $f(\lambda)$ in trigonometric terms is highlighted when expression (1.36) is written as

$$f(\lambda) = \frac{1}{2\pi} \left(\rho_0 + 2 \sum_{s=+1}^{+\infty} \rho_s \cos(s\lambda) \right). \quad (1.37)$$

Remark 5 The integrals in (1.29) and (1.35) are deterministic (classical), while the integral in (1.31) is stochastic (although with deterministic integrand; for the definition of this kind of integral see, for example, Grimmett and Stirzaker [2001]: section 9.4 for deterministic integrands and section 13.8 for stochastic ones).

Under some conditions, a stochastic process can be expressed as an infinite combination of white noise. Since the noise contains always the same information, this means that the information of the process can be encapsulated in a sequence of coefficients. The following theorem is a consequence of the *Wold decomposition* (see Wold [1938]).

Theorem 5 Any (purely indeterministic) discrete time stationary process (X_t) can be expressed in the form

$$X_t = \sum_{s=0}^{+\infty} a_s \epsilon_{t-s}, \quad t \geq 1, \quad (1.38)$$

with $a_0 = 1$, $\sum_{s=0}^{+\infty} a_s^2 < \infty$ and (ϵ_s) a white noise process. The sequences $\{a_s\}$ and (ϵ_s) are uniquely determined.

Proof. See section 10.1.5 of Priestley (1981). □

Remark 6 Due to its similarity with the finite moving average (MA) models, the expression (1.38) is known as the $MA(\infty)$ -representation of the process (X_t) .

IN PRACTICE

Let us note that in a theoretically infinite stochastic process, only a finite quantity of variables can be considered in practice,

$$(X_t, 1 \leq t \leq T) = (X_1, X_2, \dots, X_T). \quad (1.39)$$

1.3.2 Time Series Data

A time series can be interpreted as a realization of a (generating) stochastic process. For our purposes, a formal enough definition of time series is the following:

Definition 18 A time series is a sequence of numerical variables,

$$(x_t, 1 \leq t \leq T) = (x_1, \dots, x_t, \dots, x_T), \quad (1.40)$$

where the index t represents the time at which x_t is observed.

For this definition to be useful, some dependence among data is supposed. A time series is named *stationary* if its generating stochastic process is supposed stationary.

1.3.3 Statistical Inference

For the stationary time series $(x_t, 1 \leq t \leq T)$, the following inferential definitions and properties can be found, for example, in Priestley (1981).

MEAN FUNCTION

The mean function (in fact, a constant $\mu_t = \mu$) is estimated by

$$\bar{x}_T = \frac{1}{T} \sum_{t=1}^T x_t. \quad (1.41)$$

AUTOVARIANCE FUNCTION

The autocovariance function can be estimated by

$$\hat{\sigma}_\tau = \frac{1}{T} \sum_{t=1}^{T-|\tau|} (x_t - \bar{x}_T)(x_{t+|\tau|} - \bar{x}_T), \quad |\tau| = 0, 1, \dots, T-1. \quad (1.42)$$

In this expression, to obtain an unbiased estimator some authors prefer using $1/(T - |\tau|)$ instead of $1/T$ (see section 5.3.3 of Priestley [1981]); the expression (1.42) has lower mean square error.

AUTOCORRELATION FUNCTION

Now, the autocorrelation function can be estimated by

$$\hat{\rho}_\tau = \frac{\hat{\sigma}_\tau}{\hat{\sigma}_0}, \quad |\tau| = 0, 1, \dots, T-1. \quad (1.43)$$

SPECTRAL DENSITY FUNCTION

To estimate the spectral density function, the following concept is defined.

Definition 19 *The periodogram of $(x_t, 1 \leq t \leq T)$ is defined as*

$$I_T(\lambda) = \frac{1}{2\pi T} \left| \sum_{t=1}^T x_t e^{-it\lambda} \right|^2, \quad \lambda \in [-\pi, +\pi]. \quad (1.44)$$

As estimator of the spectral density function, the periodogram is asymptotically unbiased but its variance does not decrease as n increases and it is not consistent. There are several ways of constructing —basically by smoothing the periodogram— consistent estimates of the spectral density function.

SPECTRAL DISTRIBUTION FUNCTION

The integration, a sort of smoothing, provides a consistent estimate of the spectral distribution function.

Definition 20 The integrated or cumulative periodogram of $(x_t, 1 \leq t \leq T)$ is defined as $F_T(\lambda) = \int_{-\pi}^{\lambda} I_T(h)dh$, or, with a normalization,

$$F_T(\lambda) = \frac{1}{c_T} \int_{-\pi}^{\lambda} I_T(h)dh, \quad \lambda \in [-\pi, +\pi], \quad (1.45)$$

where $c_T = \int_{-\pi}^{\pi} I_T(h)dh$.

FOURIER FREQUENCIES SET

Previous definitions have been given for $\lambda \in [-\pi, +\pi]$. In practice, only a finite set of values is considered for the estimation. The choice of the set of *Fourier frequencies* provides —among others— some simplification in the algebra (see, for example, sections 6.1.3 of Priestley [1981] and 2.7 of Diggle [1990]):

$$\mathcal{S} = \left\{ \lambda_j = \frac{2\pi j}{T}, \quad j = -\left[\frac{T-1}{2}\right], \dots, -1, 0, +1, \dots, +\left[\frac{T}{2}\right] \right\}. \quad (1.46)$$

In fact, these frequencies are related to efficient ways of computing the periodogram; for example, the *fast Fourier transform*. Then,

$$I_T(\lambda_j) = \frac{1}{2\pi T} \left| \sum_{t=+1}^T x_t e^{-it\lambda_j} \right|^2, \quad \lambda_j \in \mathcal{S}, \quad (1.47)$$

and the integrated periodogram takes the form $F_T(\lambda_j) = \sum_{i=1}^j I_T(\lambda_i)$, or, in the normalized version,

$$F_T(\lambda_j) = \frac{1}{c_T} \sum_{i=1}^j I_T(\lambda_i), \quad \lambda_i \in \mathcal{S}, \quad \lambda_j \in \mathcal{S}, \quad (1.48)$$

where now $c_T = \sum_{i=1}^m I_T(\lambda_i)$, with m being the cardinal of \mathcal{S} .

POSITIVE FREQUENCIES ONLY

Both functions $f(\lambda)$ and $I_T(\lambda)$ are symmetric, so they (as well as set \mathcal{S}) could have been defined only in positive values in $[0, +\pi]$.

1.3.4 Addendum: Locally Stationary Processes

Some causes of nonstationarity —as trend, heteroscedasticity and seasonality— can be removed by applying well-known transformations. A possible approach to nonstationary processes is based on this removal. The other frequently used approach consists of supposing local stationarity

and applying the usual techniques in narrow blocks. The literature includes several definitions of “nonstationary processes”, as *locally stationary random processes*, *semi-stationary processes*, *quasi-stationary processes*, *piecewise stationary processes*, etcetera. For a stochastic process, there are many more ways of being nonstationary than stationary, and it seems that at least the local stationarity assumption is necessary. Perhaps those definitions of Priestley (1965) and Dahlhaus (1996) have been the most successful; see Dahlhaus’ paper for a comparison of both spectra. In this section, the approach of Dahlhaus is chronologically presented from several papers.

As there is a dynamic change in time, in the nonstationary framework it is not possible to separate the time and the frequency domains. The strategy of Dahlhaus started with a sort of “spectral representation” definition:

Definition 21 (*Dahlhaus [1996]*) *A sequence of stochastic processes $(X_{t,T} \ 1 \leq t \leq T, \ T \geq 1)$ is called locally stationary with transfer function A^0 and trend μ if such a representation exists*

$$X_{t,T} = \mu\left(\frac{t}{T}\right) + \int_{-\pi}^{+\pi} e^{i\lambda t} A_{t,T}^0(\lambda) d\xi(\lambda), \quad (1.49)$$

where

(i) $\xi(\lambda)$ is a stochastic process on $[-\pi, +\pi]$ with $\overline{\xi(\lambda)} = \xi(-\lambda)$ and

$$\text{cum}\{d\xi(\lambda_1), \dots, d\xi(\lambda_k)\} = \eta\left(\sum_{j=1}^k \lambda_j\right) h_k(\lambda_1, \dots, \lambda_{k-1}) d\lambda_1 \cdots d\lambda_k,$$

where $h_1 = 0$, $h_2(\lambda) = 1$, $|h_k(\lambda_1, \dots, \lambda_{k-1})| \leq \text{const}_k$ for all k , $\text{cum}\{\dots\}$ denotes the cumulant of k th order and $\eta(\lambda) = \sum_{j=-\infty}^{+\infty} \delta(\lambda + 2\pi j)$ is the period 2π extension of the Dirac delta function.

(ii) There is a constant c and a 2π -periodic function $A : [0, 1] \times \mathbb{R} \rightarrow \mathbb{C}$ with $A(u, -\lambda) = \overline{A(u, \lambda)}$ and

$$\sup_{t,\lambda} |A_{t,T}^0(\lambda) - A(t/T, \lambda)| \leq cT^{-1},$$

for all T .

$A(u, \lambda)$ and $\mu(u)$ are assumed to be continuous in u .

Remark 7 The smoothness of A in u guarantees that the process (in fact, the sequence of processes) has locally “stationary behaviour”.

Definition 22 (*Dahlhaus [1996]*) *The (time-varying) spectral density of the process (sequence of processes) is defined as:*

$$f(u, \lambda) = A(u, \lambda) \overline{A(u, \lambda)} = |A(u, \lambda)|^2. \quad (1.50)$$

For these processes, Dahlhaus (1996) also defines the local covariance of lag k at time u , and gives kernel estimates of it, as well as of the spectral density.

For the locally stationary processes, Dahlhaus and Polonik have more recently given a sort of “MA(∞)-representation”, instead of the previous spectral-type definition. Let

$$V(g) = \sup \left\{ \sum_{k=1}^m |g(x_k) - g(x_{k-1})|, \quad \text{where } 0 \leq x_0 < \dots < x_m \leq m, \quad m \in \mathbb{N} \right\} \quad (1.51)$$

be the total variation of a function g on $[0, 1]$, and for some $\kappa > 0$, let be

$$l(j) = \begin{cases} 1, & |j| \leq 1, \\ |j| \log^{1+\kappa} |j| & |j| > 1. \end{cases} \quad (1.52)$$

Definition 23 (*Dahlhaus and Polonik [2006]*) *The sequence $(X_{t,T}, 1 \leq t \leq T, T \geq 1)$ is a locally stationary process if it has the representation*

$$X_{t,T} = \sum_{j=-\infty}^{+\infty} a_{t,T}(j) \epsilon_{t-j}, \quad (1.53)$$

where the ϵ_t are identically distributed with $\mathbb{E}(\epsilon_t) \equiv 0$, $\mathbb{E}(\epsilon_s \epsilon_t) = 0$ for $s \neq t$, $\mathbb{E}(\epsilon_t^2) \equiv 1$, and satisfying the following conditions:

$$\sup_t |a_{t,T}(j)| \leq \frac{K}{\ell(j)} \quad (\text{with } K \text{ not depending on } T),$$

and there exist functions $a(\cdot, j) : (0, 1] \rightarrow \mathbb{R}$ with

$$\begin{aligned} \sup_u |a(u, j)| &\leq \frac{K}{\ell(j)}, \\ \sup_j \sum_{t=1}^T \left| a_{t,T}(j) - a\left(\frac{t}{T}, j\right) \right| &\leq K, \\ V(a(\cdot, j)) &\leq \frac{K}{\ell(j)}. \end{aligned}$$

This definition is more general than definition 21, since the parameter curves are allowed to have jumps (but bounded variation in time direction). On the other hand, this representation can easily be transformed into the previous time-varying spectral representation. The above conditions are discussed in Dahlhaus and Polonik (2009).

Definition 24 (*Dahlhaus and Polonik [2006]*) Let $(X_{t,T})$ be a locally stationary process with time-varying spectral density $f(u, \lambda)$; then

$$\sigma_{u,k} = \int_{-\pi}^{+\pi} f(u, \lambda) e^{i\lambda k} d\lambda = \sum_{j=-\infty}^{+\infty} a(u, k+j) a(u, j) \quad (1.54)$$

is the time-varying covariance of lag k at rescaled time u .

Remark 8 The integral in (1.54) is deterministic while the integral in (1.49) is stochastic.

Definition 25 (*Dahlhaus and Polonik [2006]*) Let $(X_{t,n})$ be a locally stationary process. The function

$$f(u, \lambda) = \frac{1}{2\pi} |A(u, \lambda)|^2 \quad (1.55)$$

with

$$A(u, \lambda) = \sum_{j=-\infty}^{+\infty} a(u, j) e^{-i\lambda j} \quad (1.56)$$

is the time-varying spectral density function.

1.4 Stochastic Functions

This type of data appears in many mathematical areas and has been used since long time ago. In the theory of continuous-time univariate stochastic processes, $(X(t), t \in [0, T])$, for fixed t a random variable $X(t)$ is obtained, while for a given event $\omega \in \Omega$, the *trajectories* $(X(t))$ are in a functional space L , as represented through the application

$$\begin{aligned} X(t, \cdot) : \Omega &\longrightarrow L \\ \omega &\longrightarrow X(t, \omega) \end{aligned}$$

In this case, the characteristics of the space L depends on the properties of $(X(t))$; for example, for the Wiener process the trajectories are —almost surely— continuous everywhere but differentiable nowhere.

1.4.1 Models

The following definition is based on definition 1.1 of Ferraty and Vieu (2006). Let (Ω, \mathcal{F}, P) be a probability space.

Definition 26 A stochastic function or functional random variable is a random variable, $\mathcal{X} = \mathcal{X}(\omega)$, $\omega \in \Omega$, that takes values in a functional space L .

Let $t \in [0, T]$ be the independent variable of the elements of L , then \mathcal{X} is in fact a bidimensional function, $\mathcal{X}(\omega, t)$, such that for fixed ω a deterministic function is obtained and for fixed t a random variable is obtained (this point of view, instead of the previous definition, is preferred when the dynamic behaviour is studied). For simplicity, $\mathcal{X}(\omega, t)$ will be written as \mathcal{X} , and $\mathcal{X}(t)$ when the membership to the functional space L must be highlighted. The main interest will be in the application:

$$\begin{aligned} \mathcal{X}(t, \cdot) : \Omega &\longrightarrow L \\ \omega &\longrightarrow \mathcal{X}(t, \omega) \end{aligned}$$

As a final comment, let us notice that the classical analysis differentiation makes sense only when ω is fixed. In this situation $D^i \mathcal{X}(t)$ will be denoting the i -th derivative of the real function $\mathcal{X}(t) = \mathcal{X}(\omega, t)$ of real variable t :

$$D^i \mathcal{X}(t) = \frac{d^i \mathcal{X}(t)}{dt^i}, \quad t \in [0, T], \quad (1.57)$$

even when both quantities \mathcal{X} and $D^i \mathcal{X}$ can be considered as stochastic; that is, in our framework D^i can be thought as an operator in the functional space of nonstochastic functions L or, equivalently, as an operator in the previous functional space of random functions.

Remark 9 The variable t is real and “deterministic” while, in general, ω is not. For some type of random variables, the Stochastic Calculus provides definitions of the derivative operator with respect to ω (see, for example, section 1.2 of Nualart [1995]).

For a general space L of functions, Ferraty and Vieu (2006) propose the following definitions for the functional mean.

Definition 27 The mean function of the model \mathcal{X} is defined as

$$\mu(t) = \mathbb{E}(\mathcal{X}) = \int_{\Omega} \mathcal{X}(\omega, t) dP(\omega), \quad t \in [0, T], \quad (1.58)$$

where (Ω, \mathcal{F}, P) is the probability space.

Ferraty and Vieu (2006) give also definitions for the *median function*, the *mode function* and estimators of them.

More generally, in the literature on random functions, concepts like *(auto)correlation function*, *mutual correlation function* (between two functions) or *spectral decomposition* can be found, when they exist.

IN PRACTICE

Let us notice that in this section we have not taken into account the fact that a function is continuously observed only in few cases. In practice a function consists in a pair of vectors (\mathbf{t}, \mathbf{X}) , with $\mathbf{t} = (t_1, t_2, \dots, t_T)$ and $\mathbf{X} = (X_1, X_2, \dots, X_T)$, where t_i is the time at which the X_i value is taken. Besides, \mathbf{t} could be different in each different element of a sample of functions.

1.4.2 Functional Data

Definition 28 A functional datum is an observation of a functional random variable $\mathcal{X}(\omega)$.

For functional data, a huge amount of classical theory is available. The most important for us is that on metric, normed and Hilbert spaces, and their topology and calculus. Of special interest is the theory of linear spaces, where under good conditions (mainly numerability and separability of the induced topological space) bases of functions can be used to approximate or represent the elements of the spaces. Some frequently used bases are: the basis of the monomials

$$\{1, t, t^2, t^3, \dots\}, \quad (1.59)$$

the basis of the trigonometric functions

$$\{1, \sin(t), \cos(t), \sin(2t), \cos(2t), \dots\}, \quad (1.60)$$

basis of the *spline functions*, basis of the *wavelets*, *exponential* and *power bases*, *polynomial bases*, *polygonal basis*, *step-function basis* or *constant basis*, among others.

The *filtering* techniques for functional data use some basis $\{\psi_i\}$ and work with the multivariate coefficients \mathbf{c} therefore. Note that in this case the coefficients c_i contain more information than that of mere scalars.

In the following, the functional spaces used in this thesis are presented. Given $I \subset \mathbb{R}$ compact, the set of continuous functions defined on I is denoted by

$$\mathcal{C}(I) = \{\chi : I \rightarrow \mathbb{R} \mid \chi \text{ is continuous in } t\}. \quad (1.61)$$

For $l \in \mathbb{N} \setminus \{0\}$, a more restrictive set is defined for differentiable functions,

$$\mathcal{C}^l(I) = \{\chi : I \rightarrow \mathbb{R} \mid \exists D^l \chi \text{ and } D^l \chi \in \mathcal{C}(I)\}. \quad (1.62)$$

Finally, for $m \in \mathbb{N} \setminus \{0\}$ the set of m -order integrable functions is defined as

$$\mathcal{L}^m(I) = \left\{ \chi : I \rightarrow \mathbb{R} \mid \int_I |\chi(t)|^m dt < \infty \right\}, \quad (1.63)$$

where the previous integration is usually the Lebesgue integral.

With the usual operations for functions, the three previous sets are linear spaces. Besides, it holds that

Proposition 3 *For any finite $l \in \mathbb{N} \setminus \{0\}$ and $m \in \mathbb{N} \setminus \{0\}$,*

$$\mathcal{C}^l(I) \subset \mathcal{C}(I) \subset \mathcal{L}^m(I). \quad (1.64)$$

Proof.

The first inclusion holds by definition. In the case of the second, since for a given $\chi(t) \in \mathcal{C}(I)$ a finite constant exists such that $\sup_{t \in [0, T]} |\chi(t)| \leq M$ then

$$\int_I |\chi(t)|^m dt \leq \nu(I) M^m < \infty$$

where $\nu(I)$ is the Lebesgue measure of I .

□

Now, the following statements are given for the more general spaces of integrable functions, since the other two subsets are closed —as linear subspaces— and they inherit the structures.

The set $\mathcal{L}^m(I)$ with the distance

$$d_m(\chi_1, \chi_2) = \left(\int_I |\chi_1(t) - \chi_2(t)|^m dt \right)^{1/m}, \quad \chi_k \in \mathcal{L}^m(I), \quad k = 1, 2. \quad (1.65)$$

form the metric space $(\mathcal{L}^m(I), d_m(\cdot, \cdot))$. As this metric can be defined from the norm

$$\|\chi\|_m = \left(\int_I |\chi(t)|^m dt \right)^{1/m}, \quad \chi \in \mathcal{L}^m(I), \quad (1.66)$$

the pair $(\mathcal{L}^m(I), \|\cdot\|_m)$ is a normed space. These spaces are in general numerable and separable Banach spaces, and only the case $m = 2$ is a Hilbert space, as its norm verifies the paralelogram condition and can be defined the inner product

$$\langle \chi_1, \chi_2 \rangle = \int_I |\chi_1(t) \chi_2(t)| dt, \quad \chi_k \in \mathcal{L}^2(I), \quad k = 1, 2. \quad (1.67)$$

In general, the functional space determines the allowed operations: proximity, types of convergence, continuity or geometry (for example, projection, angle or orthogonality). Expression (1.58) needs the $\mathcal{L}^1(I)$ space while expression (1.68) needs just a linear space structure (both spaces need an overlapped metric structures for the convergence). More generally, Ferraty and Vieu (2006) frequently base definitions and model adjustments on optimization problems, so they usually work in “good” Banach spaces; on the other hand, Ramsay and Silverman (2006) need the more restrictive Hilbert space structure of $(\mathcal{L}^2(I), \langle \cdot, \cdot \rangle)$, as they use the inner product to compute the coefficients of fitted models.

OUR FUNCTIONAL SPACES

In this document, the domain of the functions is compact (in \mathbb{R} with the Borel topology): in chapter 2 the interval is $I = [-\pi, +\pi]$ (or the joint of g intervals like this) while in chapter 3 it is $I = [0, T]$. Functions are continuous in t or composed of continuous parts: in chapter 2, due to the integrability of the periodogram and, in chapter 3, due to the differentiability assumptions.

As a distance measurement between two functions we have taken d_1 (defined by [1.65] with $m = 1$), so the metric or normed space of functions into which we shall be working are: the spaces $(\mathcal{C}(I), \|\cdot\|_1)$ or $(\mathcal{L}^1(I), \|\cdot\|_1)$ in chapter 2 and the more restrictive spaces $(\mathcal{C}^l(I), \|\cdot\|_1)$ in chapter 3.

1.4.3 Statistical Inference

Let $\chi_1(t), \dots, \chi_e(t), \dots, \chi_n(t)$ be a sample of *functional data*.

Definition 29 *The sample mean function of a set of functions is defined as*

$$\bar{\chi}(t) = \frac{1}{n} \sum_{e=1}^n \chi_e(t), \quad t \in [a, b] \subset \mathbb{R}, \quad (1.68)$$

and it estimates $\mu(t)$.

Ramsay and Silverman (2006) also give definitions of sample concepts like: *covariance function*, *correlation function*, *cross-covariance function* and *cross-correlation function*.

Remark 10 Although the sample mean can be computed in any linear space of functions, the ideas of approximation and convergence require some additional —stronger— analysis concept, such as norm or inner product.

Remark 11 Our proposals use the sample mean as inference tool. A worthwhile observation is that we are allowed to use this representative function due to the smoothness of our functional data, since this function does not reflect, in general, the characteristics of rough data.

1.4.4 Addendum: Functional Depth

The statistical concept of *depth* is a measurement of the “centrality” of each element inside a sample. This implies, for example, that in a set of points in $\mathbb{R}^m, m \in \mathbb{N} \setminus \{0\}$, the closer a point is to the mass center, the deeper it is. The same general idea applies to other types of data, including functions. Different definitions of depth for functions can be given. In this section we will describe and use the definitions proposed in López-Pintado and Romo (2009).

BAND DEPTH

Let $G(\chi(t)) = \{(t, \chi(t)) \mid t \in [a, b]\}$ denote the graph in \mathbb{R}^2 of a function $\chi(t)$; let $\chi_e(t), e = 1, \dots, n$, be a sample of functions; then a subset of these functions, $\chi_{e_j}(t), j = 1, \dots, m$, determines the band in \mathbb{R}^2 defined as

$$B(\chi_{e_1}(t), \dots, \chi_{e_m}(t)) = \{(t, y) \mid t \in [a, b], \min_{r=1, \dots, m} \chi_{e_r}(t) \leq y \leq \max_{r=1, \dots, m} \chi_{e_r}(t)\}. \quad (1.69)$$

For any function $\chi(t)$, the quantity

$$BD_n^{(j)}(\chi(t)) = \binom{n}{j}^{-1} \sum_{1 \leq e_1 < e_2 < \dots < e_j \leq n} \mathbb{I}\{G(\chi(t)) \subset B(\chi_{e_1}(t), \dots, \chi_{e_j}(t))\}, \quad j \geq 2, \quad (1.70)$$

expresses the proportion of bands, determined by j different curves, $\chi_{e_1}(t), \dots, \chi_{e_j}(t)$, containing the graph of $\chi(t)$ (the indicator function takes the value $\mathbb{I}\{A\} = 1$ if A occurs and $\mathbb{I}\{A\} = 0$ otherwise). The definition of depth for functional data introduced by López-Pintado and Romo (2009) states that

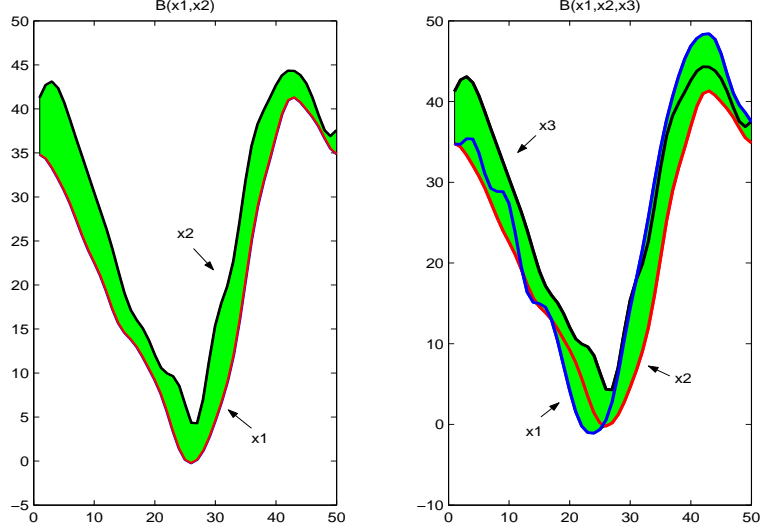
Definition 30 *For functions $\chi_e(t), e = 1, \dots, n$, the band depth of any of these curves χ with respect to the sample is*

$$BD_{n,J}(\chi(t)) = \sum_{j=2}^J BD_n^{(j)}(\chi(t)), \quad 2 \leq J \leq n. \quad (1.71)$$

If \mathcal{X} is a stochastic process and $\mathcal{X}_e(t), e = 1, \dots, n$, are copies of it, the population versions of these depth indexes are, respectively,

$$BD^{(j)}(\mathcal{X}) = P\{G(\mathcal{X}) \subset B(\mathcal{X}_{e_1}, \dots, \mathcal{X}_{e_j})\}, \quad j \geq 2, \quad (1.72)$$

Figure 1.1: Example of functions and their band



and

$$BD_J(\mathcal{X}) = \sum_{j=2}^J BD^{(j)} = \sum_{j=2}^J P\{G(\mathcal{X}) \subset B(\mathcal{X}_{e_1}, \dots, \mathcal{X}_{e_j})\}, \quad J \geq 2. \quad (1.73)$$

MODIFIED BAND DEPTH

A more flexible notion of depth is also defined in López-Pintado and Romo (2009). The indicator function in definition (1.70) is replaced by the length of the set where the function is inside the corresponding band. For any function $\chi(t)$ of $\chi_e(t)$, $e = 1, \dots, n$, and $2 \leq j \leq n$, let

$$\begin{aligned} A_j(\chi(t)) &\equiv A(\chi(t); \chi_{e_1}(t), \dots, \chi_{e_j}(t)) \\ &\equiv \{t \in [a, b] \mid \min_{r=1, \dots, j} \chi_{e_r}(t) \leq \chi(t) \leq \max_{r=1, \dots, j} \chi_{e_r}(t)\} \end{aligned} \quad (1.74)$$

be the set of points in the interval $[a, b]$ where the function $\chi(t)$ is inside the band. If ν is the Lebesgue measure on the interval $[a, b]$, $\nu_r(A_j(\chi(t))) = \nu(A_j(\chi(t)))/\nu([a, b])$ is the “proportion of time” that $\chi(t)$ is inside the band.

The quantity

$$MBD_n^{(j)}(\chi(t)) = \binom{n}{j}^{-1} \sum_{1 \leq e_1 < e_2 < \dots < e_j \leq n} \nu_r(A(\chi(t); \chi_{e_1}(t), \dots, \chi_{e_j}(t))), \quad 2 \leq j \leq n, \quad (1.75)$$

is a modified version of $BD_n^{(j)}$.

Notice that if $\chi(t)$ is always inside the band, the measure $\nu_r(A_j(\chi(t)))$ is 1 and this generalises the definition given in (1.70).

Definition 31 *The modified band depth of any of the curves $\chi(t)$ in $\chi_e(t)$, $e = 1, \dots, n$ is*

$$MBD_{n,J}(\chi(t)) = \sum_{j=2}^J MBD_n^{(j)}(\chi(t)), \quad 2 \leq J \leq n. \quad (1.76)$$

If \mathcal{X}_e , $e = 1, \dots, n$, are independent copies of the stochastic process \mathcal{X} , the population versions of these depth indexes are, respectively,

$$MBD^{(j)}(\mathcal{X}) = \mathbb{E}(\nu_r(A(\mathcal{X}; \mathcal{X}_{e_1}, \dots, \mathcal{X}_{e_j}))), \quad 2 \leq J \leq n, \quad (1.77)$$

and

$$MBD_J(\mathcal{X}) = \sum_{j=2}^J MBD^{(j)}(\mathcal{X}) = \sum_{j=2}^J \mathbb{E}(\nu_r(A(\mathcal{X}; \mathcal{X}_{e_1}, \dots, \mathcal{X}_{e_j}))), \quad 2 \leq J \leq n. \quad (1.78)$$

In chapter 2 we have used the value $J = 2$, since it is computationally fast and the modified band depth is very stable in J , providing similar center-outward order in a collection of functions (see López-Pintado and Romo [2006, 2009]).

ROBUST INFERENCE: TRIMMED MEAN

Let $\chi_{(e)}(t)$, $e = 1, \dots, n$ be a sample of functions ordered by decreasing depth, then

Definition 32 *The sample α -trimmed mean function is defined as*

$$\bar{\chi}^{\alpha}(t) = \frac{1}{n - [n\alpha]} \sum_{e=1}^{n-[n\alpha]} \chi_{(e)}(t), \quad (1.79)$$

where $[\cdot]$ is the integer part function.

A kind of “median function”, in the sense of the “deepest” curve, is obtained with the previous expression just considering $\alpha = (n - 1)/n$. Nevertheless, the α -trimmed mean is robust, like the median, and summarizes the general behaviour of the functions, like the mean. In our simulation and real data exercises, a value of $\alpha = 0.2$ is used. It means that for each group the 20% less deep data is left out.

1.5 The Classification Problem

It is frequently necessary to study a set of —abstract or real— objects in order to characterise their heterogeneity, that is, to identify the underlying structure of subsets. This task is known

as *classification*, and consists of finding some “properties” —theoretical or approximated— that characterise the differences among subsets. As the elements of each subset must be similar, some kind of “distance”, based on the characterising properties, is necessary to evaluate the proximity — between two elements, an element and a group or two groups— and, finally, to apply a “criterion” for determining the underlying structure.

When there is no a priori knowledge about the structure, the problem is termed *unsupervised classification* —or *clustering*, among other names— and even the number of subsets is usually unknown and has to be determined. If the structure of subsets is indicated by the membership of the elements in a subset (through labels), the problem is named *supervised classification*. The problems we address here are of this latter kind.

In the rest of this section, a general symbolic notation is used to locate the aforementioned concepts of “property”, “distance” and “criterion”.

1.5.1 Theoretical Rule

Let $P^{(k)}$, $k = 1, \dots, K$ be different and disjoint populations of theoretical mathematical objects, where “disjoint” means that any element belongs to only one population, and let $c^{(\cdot)}$ be a property characterising as much as possible these populations; then the classification problem consists in applying a criterion to decide to which population a new element E belongs by using the information provided by its property $c_{(E)}$. It is important to notice that the choice of $c^{(\cdot)}$ and $c_{(E)}$ usually depends on the information available or the information that we are capable to know. On the other hand, since $c^{(\cdot)}$ should be as much informative as possible, it seems natural for the election of $c_{(E)}$ to be as similar as possible to $c^{(\cdot)}$.

SEPARABLE POPULATIONS

Firstly, let us consider the ideal case in which the quantity $c^{(\cdot)}$ does characterise the populations —is different and unique for each group— and $c_{(E)}$ can contain the same information.

In the unrealistic situation in which we can in practice know the exact information of the element and of the populations, the classification is not a problem but a mere deterministic

application, $C : \cup_{k=1}^K P^{(k)} \rightarrow \{1, 2, \dots, K\}$ such that:

$$C(E) = \begin{cases} k = 1 & \text{if } c_{(E)} = c^{(1)} \\ k = 2 & \text{if } c_{(E)} = c^{(2)} \\ \vdots & \vdots \\ k = K & \text{if } c_{(E)} = c^{(K)} \end{cases} \quad (1.80)$$

The uniqueness of each $c^{(k)}$ implies that the populations are *separable*.

NONSEPARABLE POPULATIONS

In the previous case we have considered theoretically disjoint populations $P^{(k)}$, $k = 1, \dots, K$, in the sense that $c^{(\cdot)}$ is a property characterising them. Nevertheless, in most situations it happens that $c^{(\cdot)}$ provides useful information but not sufficient information for a perfect classification (an example is given below). In these cases, there is not available such a naive rule as (1.80), and some elements cannot be assigned

$$C(E) = \begin{cases} k = 1 & \text{if } c_{(E)} = c^{(1)} \\ k = 2 & \text{if } c_{(E)} = c^{(2)} \\ \vdots & \vdots \\ k = K & \text{if } c_{(E)} = c^{(K)} \\ k = \emptyset & \text{otherwise} \end{cases}, \quad (1.81)$$

where the element is assigned to the empty set if it is not equal to one, and only one, model. In this case the rule cannot assign with certainty the element E to a unique population.

1.5.2 Approximate Rule

Nevertheless, the exact theoretical information $c_{(E)}$ is usually unknown due to different causes, for example measurement and calculation errors, partial knowledge or computational aspects. In this situation, the approximation $\tilde{c}_{(E)}$ is available. If the population information is known, now the

classification rule $\tilde{C} : \cup_{k=1}^K P^{(k)} \rightarrow \{1, 2, \dots, K, \emptyset\}$ has an uncertain component:

$$\tilde{C}(E) = \begin{cases} k = 1 & \text{if } \tilde{c}_{(E)} \approx c^{(1)} \\ k = 2 & \text{if } \tilde{c}_{(E)} \approx c^{(2)} \\ \vdots & \vdots \\ k = K & \text{if } \tilde{c}_{(E)} \approx c^{(K)} \\ k = \emptyset & \text{otherwise} \end{cases} \quad (1.82)$$

where the element is assigned to the empty set if it is not similar to one, and only one, model.

1.5.3 Sample Rule

Since in general the theoretical populations are defined to model the reality, the exact information $c^{(k)}$ is not known, and the previous rule has to be substituted by the empirical version, $\hat{C} : \cup_{k=1}^K P^{(k)} \rightarrow \{1, 2, \dots, K, \emptyset\}$.

$$\hat{C}(E) = \begin{cases} k = 1 & \text{if } \tilde{c}_{(E)} \approx \hat{c}^{(1)} \\ k = 2 & \text{if } \tilde{c}_{(E)} \approx \hat{c}^{(2)} \\ \vdots & \vdots \\ k = K & \text{if } \tilde{c}_{(E)} \approx \hat{c}^{(K)} \\ k = \emptyset & \text{otherwise} \end{cases} \quad (1.83)$$

where the estimates $\hat{c}^{(k)}$ are constructed from samples of the populations. In practice, the element E is assigned to the closest population. Again the element is assigned to the empty set if it is not similar to one, and only one, model. In this case the classification rule is still less certain than the previous $\tilde{C}(\cdot)$ and a stochastic component is added by the use of samples. If $E_1^{(k)}, \dots, E_{n_k}^{(k)}$ is a sample of the k -th population, usually $\hat{c}^{(k)}$ is such that

$$\hat{c}^{(k)} = \frac{1}{n_k} \sum_{e=1}^{n_k} \tilde{c}_{E_e^{(k)}}, \quad (1.84)$$

where the term $\tilde{c}_{E_e^{(k)}}$, of the e -th element $E_e^{(k)}$ of the k -th sample, has both approximative and stochastic characters. The average (1.84) is the sample version of the population quantity $c^{(k)} = \mathbb{E}(c_{E^{(k)}})$, where $E^{(k)}$ follows the theoretical functional distribution of the k -th population and $c_{E^{(k)}}$ is its property.

Nonetheless, before using expression (1.84) it is necessary to be sure that such a mix of information (the sample mean) makes sense and is useful. For example, if the property of the populations were the *range of values* (e.g. of continuous random variables), averaging would be a quite

inappropriate way of mixing the information, a more proper estimator would be

$$\hat{c}^{(k)} = [\min_{e=1, \dots, n_k} \{\tilde{c}_{E_e^{(k)}}\}, \max_{e=1, \dots, n_k} \{\tilde{c}_{E_e^{(k)}}\}] \quad (1.85)$$

1.5.4 Stochastic Rule

So far we have relegated some difficult or uncomfortable decisions —e.g. nonseparable populations— to the “otherwise” case of the classification rules. On the other hand, the rules tried to assign a new element with certainty, instead of with partial uncertainty. Finally, the stochastic character of the rule (1.83) has been mentioned only implicitly, while the estimators are random quantities. The previous enuntiations of the rules highlighted the similarity of the properties and the certainty in the classification.

For several reasons, and sometimes for necessity, it is usually more convenient an enuntiation in terms of probabilities.

$$\hat{C}(E) = \begin{cases} k = 1 & \text{with probability } p_1(\tilde{c}_{(E)}, \hat{c}^{(1)}, \hat{c}^{(2)}, \dots, \hat{c}^{(K)}) \\ k = 2 & \text{with probability } p_2(\tilde{c}_{(E)}, \hat{c}^{(1)}, \hat{c}^{(2)}, \dots, \hat{c}^{(K)}) \\ \vdots & \vdots \\ k = K & \text{with probability } p_K(\tilde{c}_{(E)}, \hat{c}^{(1)}, \hat{c}^{(2)}, \dots, \hat{c}^{(K)}) \\ k = \emptyset & \text{with probability } p_0(\tilde{c}_{(E)}, \hat{c}^{(1)}, \hat{c}^{(2)}, \dots, \hat{c}^{(K)}) \end{cases}, \quad (1.86)$$

with $\sum_{k=0}^K p_k(\tilde{c}_{(E)}, \hat{c}^{(1)}, \hat{c}^{(2)}, \dots, \hat{c}^{(K)}) = 1$. Classifying an element E in terms of proper probabilities is more informative than not classifying it. The similarity and distances can be used to compute these probabilities, but this is a complex task out of the scope of this discussion. While the election of the probabilities may be simple for separable populations, how to compute them may be quite complex for nonseparable populations. Finally, let us notice that it is easy to obtain the rule (1.83) as a particular case of this rule (1.86).

1.5.5 Asymptotic Rule

The encouraging news is that if the property $c^{(\cdot)}$, the approximations $\tilde{c}_{(\cdot)}$ and $\tilde{d}(\cdot, \cdot)$, and the estimates $\hat{c}^{(k)}$ are properly defined, the classification rule $\hat{C}(\cdot)$ tends to provide the same classification as the theoretic rule $C(\cdot)$:

$$\hat{C} \longrightarrow \tilde{C} \longrightarrow C \quad (1.87)$$

This means that the rule \hat{C} converges to C in the sense that asymptotically \hat{C} tends to classify the element E as C would do. Notice that the election of $\tilde{c}_{(\cdot)}$ and $\hat{c}^{(k)}$ should depend on the previous election of $c_{(E)}$ and $c^{(\cdot)}$.

Both relations $=$ and \approx can be expressed by the distance $\tilde{d}(\cdot, \cdot)$, and it can be written informally:

$$\tilde{d}(c_{(E)}, c^{(k)}) \leq \tilde{d}(c_{(E)}, \tilde{c}_{(E)}) + \tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)}) + \tilde{d}(\hat{c}^{(k)}, c^{(k)}),$$

and, analogously,

$$\tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)}) \leq \tilde{d}(\tilde{c}_{(E)}, c_{(E)}) + \tilde{d}(c_{(E)}, c^{(k)}) + \tilde{d}(c^{(k)}, \hat{c}^{(k)}).$$

Thus, when $\tilde{d}(\tilde{c}_{(E)}, c_{(E)}) \rightarrow 0$ and $\tilde{d}(\hat{c}^{(k)}, c^{(k)}) \rightarrow 0$ it is concluded that $\tilde{d}(c_{(E)}, c^{(k)}) \leq \lim\{\tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)})\}$ and that $\lim\{\tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)})\} \leq \tilde{d}(c_{(E)}, c^{(k)})$, so

$$\lim\{\tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)})\} = \tilde{d}(c_{(E)}, c^{(k)}). \quad (1.88)$$

Note that several populations can be theoretically separable (the models), but at the same time their respective samples (the data) may not be separable: perhaps there are elements that could have been generated by the different models and the information available from the property does not characterise the groups. Asymptotically, what usually happens is that the unlimited amount of information in the samples —of the supervised classification framework— usually allows the separability of the samples. Then, in spite of the uncertainty and random components, a classification could be “asymptotically perfect”, where *perfect* must be understood in the sense that, for separable populations,

$$\left\{ \begin{array}{ll} P(\hat{C}(E) = 1) = 1 & \text{if } E \in P^{(1)} \\ P(\hat{C}(E) = 2) = 1 & \text{if } E \in P^{(2)} \\ \vdots & \vdots \\ P(\hat{C}(E) = K) = 1 & \text{if } E \in P^{(K)} \end{array} \right., \quad (1.89)$$

and, for nonseparable populations,

$$\left\{ \begin{array}{ll} P(\hat{C}(E) = 1) = 1 & \text{if } E \in P^{(1)} \cap \{\cup_{k \neq 1} P^{(k)}\}^c \\ P(\hat{C}(E) = 2) = 1 & \text{if } E \in P^{(2)} \cap \{\cup_{k \neq 2} P^{(k)}\}^c \\ \vdots & \vdots \\ P(\hat{C}(E) = K) = 1 & \text{if } E \in P^{(K)} \cap \{\cup_{k \neq K} P^{(k)}\}^c \\ P(\hat{C}(E) = 1) = p_{12}^{(1)} & \text{if } E \in P^{(1)} \cap P^{(2)} \cap \{\cup_{k \neq 1,2} P^{(k)}\}^c \\ P(\hat{C}(E) = 2) = p_{12}^{(2)} & \text{if } E \in P^{(1)} \cap P^{(2)} \cap \{\cup_{k \neq 1,2} P^{(k)}\}^c \\ \vdots & \vdots \end{array} \right., \quad (1.90)$$

where the superindex c denotes the complementary set and, in the place of the second set of vertical points, all possible intersections among the populations must be considered (only the case of the intersection of the two first populations have been included). The election of the probabilities can follow lots of different criteria, e.g. the probabilities being proportional to the likelihood of having been generated by the model of each class.

1.5.6 Example

To illustrate the previous frameworks, let us consider random variables following discrete uniform laws with different and disjoint (nonoverlapping) set of values; these populations are different and disjoint (any variable follows only one of the laws). The groups are characterised by the property *set of values*. For a new variable the most similar information to the set of values is the *single value* of the variable, that is also characterising information. Let us measure the proximity of a single value to a set of values through the minimum of the absolute values of the distances from the single value to the values of the set. Under the knowledge of all the previous information, the criterion (1.80) can be applied without problems. Nevertheless, several complexities arise immediately. On the one hand, we have some limitation in managing numbers and operations, since we can work with a high but limited level of precision; this implies that the criterion (1.82) must be considered instead of (1.80). On the other hand, when the property *set of values* of the populations is not known, it is necessary to use some estimator of it (notice that in this case an expression like (1.84) makes no sense). If the estimator of the *set of values* is a record of the values of a sample and the approximations are good enough, asymptotically there will be enough information to classify as stated in (1.89).

On the other hand, if the sets of values of these previous discrete uniform laws are not disjoint, that is, they overlap, then we are in the case of nonseparable populations. The main difference is that the property *single value* of the variable is not characterising, and some rule of the form (1.81) must be considered at the beginning. The technical problems mentioned in the previous paragraph would also remain in this case. Finally, asymptotically and under good conditions the classification would take the form (1.90).

1.5.7 Our Classification Frameworks

MINIMIZING THE DISTANCE

The role of the “similarity of the property” has been highlighted in the previous classification criteria, but this similarity (equality and approximation) has to be measured with a distance. The theoretical criteria are, in this case, equivalent to

$$k = \operatorname{argmin}_{\{1, \dots, K\}} \{d(c_{(E)}, c^{(k)})\}, \quad (1.91)$$

since all the values $d(c_{(E)}, c^{(k)})$, $k = 1, \dots, K$ are constant but one. Then, it seems natural assigning a new element to the population minimizing the distance to the group representative; that is, the approximate and sample criteria could be substituted, respectively, by

$$k = \operatorname{argmin}_{\{1, \dots, K\}} \left\{ \tilde{d}(\tilde{c}_{(E)}, c^{(k)}) \right\} \quad (1.92)$$

and

$$k = \operatorname{argmin}_{\{1, \dots, K\}} \left\{ \tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)}) \right\}, \quad (1.93)$$

where $\tilde{d}(\cdot, \cdot)$ denotes the fact that, in general, the exact theoretical distance cannot be computed.

Now, without taking into account the discretization —of curves, integrals, etc.— for computational necessity, the previous general objects will take the following mathematical meaning for the classification methods proposed in chapters 2 and 3:

TIME SERIES METHOD

$c^{(k)}$	Curve formed from the spectral distribution function models
$c_{(E)}$	Curve formed in the same way for the time series E
$d(\cdot, \cdot)$	Natural distance of the functional space \mathcal{L}^1
$\hat{c}^{(k)}$	Empirical curve estimate formed from integrated periodograms
$\tilde{c}_{(E)}$	Curve formed from the integrated periodogram of the series E
$\tilde{d}(\cdot, \cdot)$	Approximation of the previous distance
$\hat{C}(E)$	Criterion such that $k = \operatorname{argmin}_{\{1, 2\}} \left\{ \tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)}) \right\}$.

FUNCTIONAL DATA METHOD

$c^{(k)}$	Univariate variable $Y^{(\cdot)}$ formed from the functional models
$c_{(E)}$	Univariate variable formed in the same way for the function E
$d(\cdot, \cdot)$	Natural distance of \mathbb{R}
$\hat{c}^{(k)}$	Empirical value $y^{(k)}$ formed from samples
$\tilde{c}_{(E)}$	Value $y_{(E)}$ formed from the curve E
$\tilde{d}(\cdot, \cdot)$	Approximation of the previous distance
$\hat{C}(E)$	Criterion such that $k = \operatorname{argmin}_{\{1,2\}} \left\{ \tilde{d}(\tilde{c}_{(E)}, \hat{c}^{(k)}) \right\}$.

Chapter 2

Time Series Classification

Summary: We propose using the integrated periodogram to classify time series. The method assigns a new element to the group minimizing the distance from the element integrated periodogram to the group mean of integrated periodograms. Local computation of these periodograms allows the application of this approach to nonstationary time series. Since the integrated periodograms are functional data, we apply depth-based functional techniques to make the classification robust. The method provides small error rates with both simulated and real data, and shows good computational behaviour.

Key words: time series, classification, integrated periodogram, depth.

2.1 Introduction

Classification of time series is a statistical subject with many applications. Time series can be studied from both time and frequency domains; while the former uses position or time as an index, the latter involves the frequency. With short, stationary series, a time domain approach based on usual multivariate techniques can be applied. Nevertheless, a frequency domain approach is more appropriate with long time series because it provides a reduction of the dimension. Moreover, frequency domain is particularly important for nonstationary series (Huang et al. [2004]). There are many studies on classification methods for stationary processes in both domains (see references in Chapter 7 of Taniguchi and Kakizawa [2000]). Several authors have addressed the discrimination

analysis of nonstationary time series: Hastie et al. (1995), Shumway (2003), Huang et al. (2004), Hirukawa (2004), Sakiyama and Taniguchi (2004), Chandler and Polonik (2006) and Maharaj and Alonso (2007), among others. Caiado et al. (2006) define a measure based on the normalized periodogram and use it for both clustering and classifying between stationary and nonstationary time series. Our procedure uses the integrated periodograms to classify time series and therefore is a frequency domain approach.

Since the integrated periodogram can be seen as a function, we shall use specific techniques for functional data analysis. There are several studies on the statistical analysis of functional data and, particularly, on their classification. For example, a penalized discriminant analysis is proposed in Hastie et al. (1995); it is adequate for situations with many highly correlated predictors, like those obtained by discretizing a function. Nonparametric tools to classify a set of curves have been introduced in Ferraty and Vieu (2003), where authors calculate the posterior probability of belonging to a given class of functions by using a consistent kernel estimator. A new method for extending classical linear discriminant analysis to functional data has been analyzed in James and Hastie (2001); this technique is particularly useful when only fragments of the curves are observed. The problem of unsupervised classification or clustering of curves is addressed in James and Sugar (2003), who elaborate a flexible model-based approach for clustering functional data; it is effective when the observations are sparse, irregularly spaced or occur at different time points for each subject. In Abraham et al. (2003) unsupervised clustering of functions is considered; they fit the data by B-splines and partition is done over the estimated model coefficients using a k-means algorithm. In a related problem, Hall et al. (2001) explore a functional data-analytic approach to performing signal discrimination. Many of these procedures are, however, highly sensitive to outliers. A natural idea for classifying functions is to minimize the distance between the new curve and a reference one from the group. The approach presented in this chapter follows this idea. As a reference function of each group, we shall take the mean of the integrated periodograms of its elements. Later, this curve will be substituted by a more robust representative.

The notion of statistical *depth* was first introduced for multivariate data. It measures the “centrality” or “outlyingness” of an observation within a set of data (or with respect to a probability distribution), providing a criterion for ordering observations from center-outward. This idea of depth has been extended to functional data in several papers (see, e.g., Fraiman and Muniz [2001] and López-Pintado and Romo [2009]). Moreover, López-Pintado and Romo (2006) have

used this concept to classify curves. Since robustness is an interesting feature of the statistical methods based on depth, we have applied these ideas to add robustness to our time series classification procedure. Their method considers the α -trimmed mean as a reference curve of each group, which is defined as the average of the $1 - \alpha$ proportion of deepest curves from the sample; in other words, it leaves $100\alpha\%$ of the least representative data curves out.

The chapter is organized as follows. In section 2, we include some definitions and describe the classification algorithm. In section 3, we explain how depth can be used to make the method robust. The next two sections, 4 and 5, show the behaviour of the procedure with simulated and real data, respectively. A brief summary of conclusions is given in section 6.

2.2 The Classification Method

A first and important step in our classification proposal is to transform the time series problem into a functional data problem by considering the integrated periodogram of each time series.

2.2.1 The Integrated Periodogram

The Fourier transform of the correlation function of an absolutely summable stochastic process is known as *spectral density* or *spectrum*; its integration provides the *spectral distribution function* or *cumulative spectrum*. Let (X_t) be a stationary process with autocovariance function $\sigma_s = \text{cov}(X_t, X_{t-s})$ satisfying $\sum_{s=-\infty}^{+\infty} |\sigma_s| < +\infty$. Then the spectral density can be expressed in terms of the autocorrelation as $f(\lambda) = \sum_{s=-\infty}^{+\infty} \rho_s \exp(-2\pi i s \lambda)$, and it holds that $\rho_s = \int_{-1/2}^{+1/2} \exp(2\pi i s \lambda) dF(\lambda)$, where $F(\cdot)$ is the spectral distribution function.

The *periodogram* is the sample version of the population concept of spectral density, and it expresses the contribution of the frequencies to the variance of a series. Let $(x_t^{(k)}) = (x_1^{(k)}, \dots, x_T^{(k)})$ be a time series of the k -th population; the periodogram $I_T^{(k)}$ is obtained as indicated in (1.47):

$$I_T^{(k)}(\lambda_j) = \frac{1}{2\pi T} \left| \sum_{t=1}^T x_t^{(k)} e^{-it\lambda_j} \right|^2, \quad \lambda_j \in \mathcal{S}. \quad (2.1)$$

Its cumulative version is the integrated periodogram $F_T^{(k)}$ computed as indicated in (1.48), that is,

$$F_T^{(k)}(\lambda_j) = \frac{1}{c_{T,k}} \sum_{i=1}^j I_T^{(k)}(\lambda_i), \quad \lambda_j \in \mathcal{S}, \quad \lambda_i \in \mathcal{S}, \quad (2.2)$$

where $c_{T,k} = \sum_{i=1}^m I_T^{(k)}(\lambda_i)$. The normalized version of the cumulative periodogram takes into account the shape of the curves more than the nonnormalized version, which considers also the scale.

In our case, we propose using the normalized version when the graphs of the functions of the different groups tend to intersect and there is no clear scale pattern, and using the nonnormalized one when the graphs do not tend to intersect.

Some of the advantages of using the integrated periodogram are: it is a nondecreasing and quite smooth curve; it has good asymptotic properties (for example, while the periodogram is an asymptotically unbiased but inconsistent estimator of spectral density, the integrated periodogram is a consistent estimator of spectral distribution); although, in practice, for stationary processes the integrated spectrum is usually estimated via the estimation of the spectrum, from a theoretical point of view, spectral distribution always exists, whereas spectral density exists only under absolutely continuous distributions (see theorem 2); finally, from a theoretical point of view, the integrated spectrum completely determines the stochastic processes.

Since the periodogram is defined only for stationary series, in order to be able to classify nonstationary time series, we shall consider locally stationary series. With this assumption we can split them into blocks, compute the integrated periodogram of each block and merge these periodograms into a final curve; hence, we approximate the locally stationary processes by piecewise stationary processes. In figure 2.1(b), we illustrate our blockwise spectral distribution estimation of the locally stationary process spectrum. It is worth mentioning that there are two opposite effects as a consequence of splitting: one is that the narrower the blocks are, the closer we are to the locally stationary assumption; the other one is that when the length of the blocks decreases, the quality of the integrated periodogram as an estimator of the integrated spectrum also decreases.

2.2.2 Classifying Functions

When functions need to be classified, a possible criterion is to assign them to the group minimizing some distance from the new data to the group. In our context this criterion means that we classify new series in the group minimizing the distance between the integrated periodogram of the series and a reference curve from the group. As a reference function of each group, we take the mean of its elements, as it summarizes the general behaviour of the sample. Let $\chi_e^{(k)}(\lambda)$ be the joint integrated periodograms of the blocks for the e -th series —out of n_k — in group k . The mean is

defined as:

$$\bar{\chi}^{(k)}(\lambda) = \frac{1}{n_k} \sum_{e=1}^{n_k} \chi_e^{(k)}(\lambda). \quad (2.3)$$

As a distance measurement between two functions we have taken the distance given by expression (1.65) with $m = 1$. Notice that the functions we are working with, that is, the integrated periodograms, belong to the $\mathcal{L}^1[-\pi, +\pi]$ space. Some other distance could be considered, and in general there is no “best” one. For example, with the usual distance of \mathcal{L}^2 , big differences between functions would be highlighted and so would be the corresponding values of the independent variable (frequency).

2.2.3 The Algorithm

We can establish the classification algorithm with the above definitions:

ALGORITHM 1

Let $(x_t)_e^{(k)}$, $e = 1, \dots, n_k$, be a sample containing n_k time series from population $P^{(k)}$, for $k = 1, 2$; the classification method includes the following steps:

1. **From time series to functions.** To this end, each time series is split into G blocks, then a curve associated with each series is constructed by merging the integrated periodograms of the blocks. Concretely, consider $\{\chi_1^{(k)}(\lambda), \dots, \chi_{n_k}^{(k)}(\lambda)\}$, $k = 1, 2$, where $\chi_e^{(k)}(\lambda) = (F_{1,e}^{(k)}(\lambda) \dots F_{G,e}^{(k)}(\lambda))$ and $F_{g,e}^{(k)}(\lambda)$ is the integrated periodogram of the g -th block of the e -th series of sample k .
2. **The reference functions $\mathcal{R}^{(k)}$.** Calculate the sample mean of each group of curves: $\mathcal{R}^{(k)}(\lambda) = \bar{\chi}^{(k)}(\lambda)$, $k = 1, 2$.
3. **The allocation of new series.** Let $\chi(\lambda)$ be the associated curve of a new series (x_t) , that is $\chi(\lambda) = (F_1(\lambda) \dots F_G(\lambda))$; then (x_t) is classified as

$$\begin{cases} k = 1 & \text{if } d(\chi(\lambda), \mathcal{R}^{(1)}(\lambda)) < d(\chi(\lambda), \mathcal{R}^{(2)}(\lambda)) \\ k = 2 & \text{otherwise} \end{cases} \quad (2.4)$$

Remark 12 An important point of our approach is that it can be interpreted as the fit to locally stationary processes with piecewise stationary processes (see figure 2.1).

Remark 13 To apply the algorithm to stationary series, G can be set equal to one. We have used a dyadic splitting of the series into blocks in the simulation and real data computations, that is, $G = 2^p, p = 0, 1, \dots$; but the implementation with blocks of different lengths, as could be suggested by visual inspection of data, is also possible.

Remark 14 The same methodology we propose in this thesis could be implemented using different classification criterion between curves, reference function of each group (as we do in the following section) or distance between curves.

2.3 Robust Version

Our classification method depends on the group reference curve to which the distance is measured. The mean of a set of functions is not robust to the presence of outliers. Then robustness can be added to the classification procedure by using a robust reference curve. Instead of considering the mean of the integrated periodograms of all the elements of the group, we shall consider the α -trimmed mean, where only the deepest elements are averaged. The trim adds robustness by making the reference curve more resistant to the presence of outliers. In this section, we describe the concept of depth for functional data given by López-Pintado and Romo (2009). Then we propose a robust version of our classification algorithm.

2.3.1 New Reference Function

In order to add robustness to the algorithm presented in the previous section, now we take the group α -trimmed mean of its elements

$$\bar{\chi}^{(\alpha)}_{(k)}(\lambda) = \frac{1}{n_k - [n_k \alpha]} \sum_{e=1}^{n_k - [n_k \alpha]} \chi_{(e)}^{(k)}(\lambda), \quad (2.5)$$

where $[\cdot]$ is the integer part function and $\chi_{(e)}^{(k)}(\lambda)$, $e = 1, \dots, n_k$ is the k -th sample of functions ordered by decreasing depth.

2.3.2 New Algorithm

With this little —but essential— difference, the algorithm hardly changes. In step 2, the group α -trimmed mean is taken instead of the group mean; thus, the distance from the series to the class is now measured using a different reference curve of the group.

ALGORITHM 2

Let $(x_t)_e^{(k)}$, $e = 1, \dots, n_k$, be a sample containing n_k time series from population $P^{(k)}$, for $k = 1, 2$; the classification method includes the following steps:

1. **From time series to functions.** To this end, each time series is split into G blocks, then a curve associated with each series is constructed by merging the integrated periodograms of the blocks. Concretely, consider $\{\chi_1^{(k)}(\lambda), \dots, \chi_{n_k}^{(k)}(\lambda)\}$, $k = 1, 2$, where $\chi_e^{(k)}(\lambda) = (F_{1,e}^{(k)}(\lambda) \dots F_{G,e}^{(k)}(\lambda))$ and $F_{g,e}^{(k)}(\lambda)$ is the integrated periodogram of the g -th block of the e -th series of sample k .
2. **The reference functions $\mathcal{R}^{(k)}$.** Calculate the sample mean of each group of curves: $\mathcal{R}^{(k)}(\lambda) = \bar{\chi}^{(k)}(\lambda)$, $k = 1, 2$.
3. **The allocation of new series.** Let $\chi(\lambda)$ be the associated curve of a new series (x_t) , that is $\chi(\lambda) = (F_1(\lambda) \dots F_G(\lambda))$; then (x_t) is classified as

$$\begin{cases} k = 1 & \text{if } d(\chi(\lambda), \mathcal{R}^{(1)}(\lambda)) < d(\chi(\lambda), \mathcal{R}^{(2)}(\lambda)) \\ k = 2 & \text{otherwise} \end{cases} \quad (2.6)$$

Remark 15 The same algorithm could be implemented using a different functional depth.

Remark 16 The previous classification criteria (algorithms 1 and 2) can both be expressed as

$$k = \operatorname{argmin}_{\{1,2\}} \{d(\chi(\lambda), \mathcal{R}^{(k)}(\lambda))\}. \quad (2.7)$$

2.4 Simulation Results

In this section, we evaluate —based on simulation studies— the two algorithms we have introduced and, as a reference, the method proposed in Huang et al. (2004). The results obtained with algorithm 1, algorithm 2 and Huang et al. (2004) are denoted by DbC, DbC- α and SLEXbC, respectively. Ombao et al. (2001) introduced the SLEX (smooth localized complex exponentials) model of a nonstationary random process, which is based on a set of Fourier-type bases that are at the same time orthogonal and localized in both time and frequency domains. The method of Huang et al. (2004) uses SLEX for classification of nonstationary time series. In a first step, they select from SLEX a basis explaining the difference between the classes of time series as well as possible. In a second step, they construct a discriminant criterion that

is related to the SLEX spectra of the different classes: a time series is assigned to the class minimizing the Kullback-Leibler divergence between the estimated spectrum and the spectrum of the class. For the SLEXbC method, we have used an implementation provided by the authors (<http://www.stat.uiuc.edu/~ombao/research.html>). To select the parameters for this method, we have carried out a small optimization for each simulation exercise and the results were similar to the values recommended by the authors.

We have used the same models as the ones proposed in Huang et al. (2004). For each model, we run the following steps 1000 times. We generate training and test sets of each class. Training sets have the same sizes (sample size and series length) as the ones used in Huang et al. (2004). The test sets always contain 10 series of length determined in each particular simulation exercise. The performance of the different methods are based on exactly the same simulated data.

Simulation Exercise 1. We compare an autoregressive process of order one $X_t^{(1)}$ with a Gaussian white noise $X_t^{(2)}$:

$$X_t^{(1)} = \phi X_{t-1}^{(1)} + \epsilon_t^{(1)} \quad t = 1, \dots, T$$

$$X_t^{(2)} = \epsilon_t^{(2)} \quad t = 1, \dots, T$$

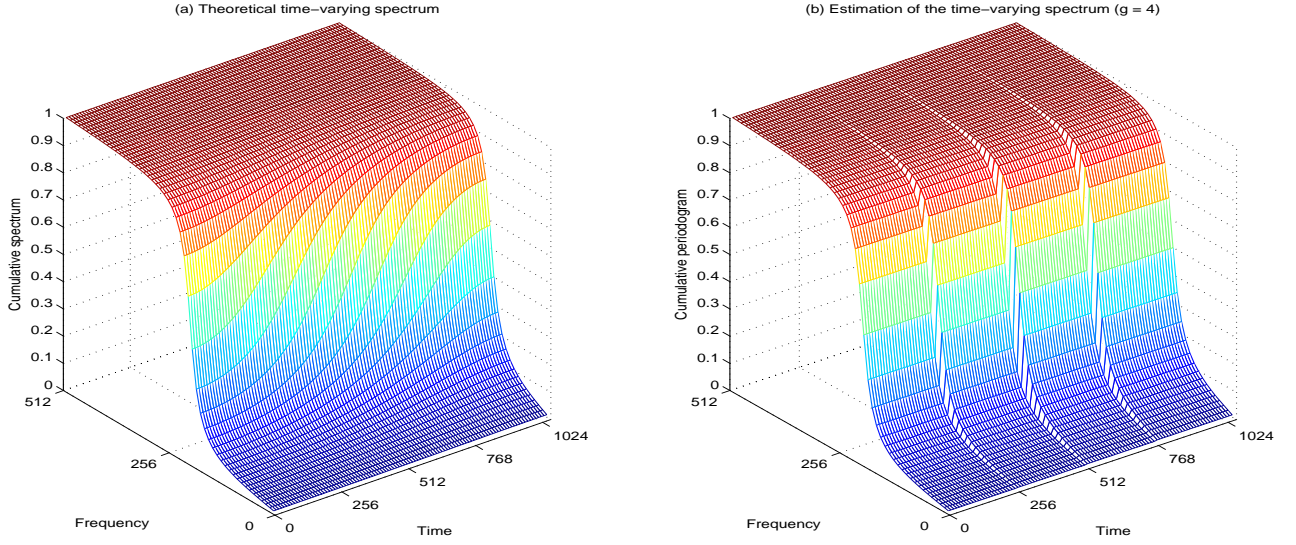
where ϵ_t are i.i.d. $N(0, 1)$, independently generated for the two models. Each training data set has $n = 8$ series of length $T = 1024$. Six comparisons have been run, with the parameter ϕ of the AR(1) model taking the values $-0.5, -0.3, -0.1, +0.1, +0.3$ and $+0.5$. Series are stationary in this exercise.

Simulation Exercise 2. We compare two processes, half of each model is white noise and half is an autoregressive process of order one. The value of the AR(1) parameter is -0.1 in the first class and $+0.1$ in the second class:

$$\begin{aligned} X_t^{(1)} &= \epsilon_t^{(1)} & \text{if } t = 1, \dots, T/2 \\ X_t^{(1)} &= -0.1X_{t-1}^{(1)} + \epsilon_t^{(1)} & \text{if } t = T/2 + 1, \dots, T \end{aligned}$$

$$\begin{aligned} X_t^{(2)} &= \epsilon_t^{(2)} & \text{if } t = 1, \dots, T/2 \\ X_t^{(2)} &= +0.1X_{t-1}^{(2)} + \epsilon_t^{(2)} & \text{if } t = T/2 + 1, \dots, T \end{aligned}$$

Figure 2.1: Time-varying autoregressive model with $\tau = 0.4$



Different combinations of training sample sizes, $n = 8$ and 16 , and series lengths, $T = 512$, 1024 and 2048 , are considered. In this exercise, the series are piecewise stationary, although the series themselves are not stationary.

Simulation Exercise 3. In this exercise, the stochastic models of both classes are slowly time-varying second-order autoregressive processes:

$$X_t^{(1)} = a_{t;0.5}X_{t-1}^{(1)} - 0.81X_{t-2}^{(1)} + \epsilon_t^{(1)} \quad t = 1, \dots, T$$

$$X_t^{(2)} = a_{t;\tau}X_{t-1}^{(2)} - 0.81X_{t-2}^{(2)} + \epsilon_t^{(2)} \quad t = 1, \dots, T$$

with $a_{t;\tau} = 0.8[1 - \tau \cos(\pi t/1024)]$, where τ is a parameter. Each training data set has $n = 10$ series of length $T = 1024$. Three comparisons have been made, the first class always having the parameter $\tau = 0.5$, and the second class having the values $\tau = 0.4$, 0.3 and 0.2 , respectively. Notice that a coefficient of the autoregressive structure is not fixed but it varies in time; therefore, the processes are not stationary. We have also proved that, for these values of τ and any value of t , the characteristic polynomial of the autoregressive process has roots outside the unit circle. See figure 2.1(a) for an example of the integrated spectrum corresponding to these processes.

In order to test the robustness of our classification procedure and the SLEXbC method, we perform additional experiments where the training set is contaminated with an outlier. In all

cases we contaminate the $P^{(1)}$ population by changing one series for another following a different model. We consider three levels of contamination: one type of weak contamination (A) and two strong contaminations (B and C).

Contamination A. For exercise 1, we replace the autoregressive structure by a moving average structure, that is, generate an MA(1) instead of an AR(1) model, with the MA parameter equal to the AR parameter. For exercise 2, we make the same substitution of structures in the autoregressive half of one series of a class (the other half remains as a white noise). For exercise 3, we contaminate the set of slowly time-varying autoregressives of parameter $+0.5$ with a series of the same model but with parameter value $+0.2$.

Contamination B. This contamination consists of using a parameter value of $\phi = -0.9$ in exercises 1 and 2 and $\tau = -0.9$ in exercise 3 for one time series instead of the correct value. Therefore, we always use the correct model except for one time series where the parameter value is mistaken.

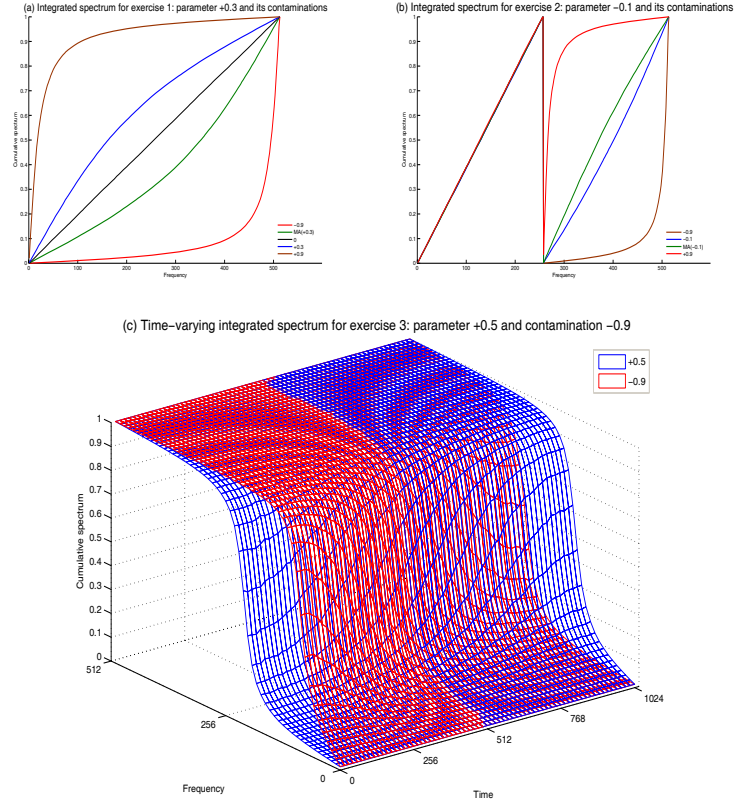
Contamination C. The same as contamination B but using a value $+0.9$ instead of the value -0.9 .

In figures 2.2(a) and 2.2(b), we illustrate the three contaminations for the first two exercises with specific parameter values. Figure 2.2(c) shows contamination B for the third exercise.

The error rate estimates for the first simulation exercise are presented in table 2.1, for the second simulation experiment in tables 2.2, 2.3, 2.4 and 2.5, and for the third simulation experiment in tables 2.6, 2.7, 2.8 and 2.9. Each cell includes the mean and the standard error (in parentheses) of the error rates based on 1000 runs.

Tables 2.10, 2.11 and 2.12 provide the estimates of the computation times of the different classification methods using the simulation exercises previously described. In these tables, each cell contains the average time in seconds to compute 1000 runs. The time is measured from the instant the series are input into the algorithm until the moment the method gives the error rate. The time required to generate the training and test time series is not included in the computation;

Figure 2.2: Examples of contaminations for the three simulation experiments



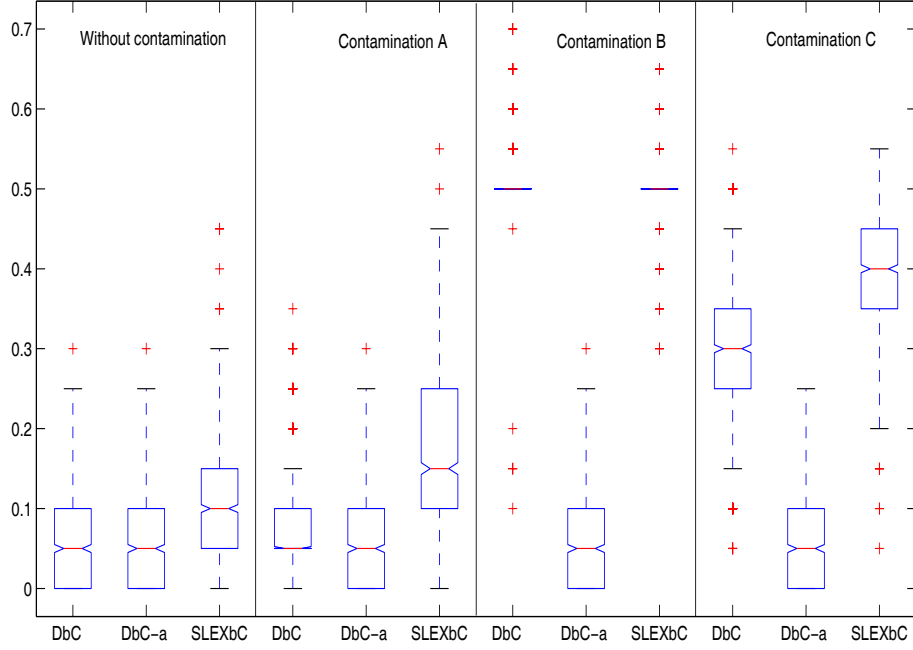
however, for our method, the computation does include the construction of functional data from the time series and the calculation of depth inside groups. Simulation exercises have been run on a personal computer with an AMD Athlon(tm) 64 Processor 3200+, 2.01GHz and 2.00Gb of RAM memory.

For all tables, we use the following notation: DbC (from *depth-based classification*) for algorithm 1, DbC- α for algorithm 2 and SLEXbC for the method of Huang et al. (2004). If a number follows DbC or DbC- α , this represents the number of blocks (k) into which the series are split. The digits in bold correspond to the minimum misclassification rates (when there is at least one value different from zero).

COMMENTS ON ERROR RATES

Table 2.1 shows the estimates of the misclassification rates for the first simulation exercise. We can observe that when there is no contamination, DbC and DbC- α provide similar error rates, and they are of lower value (approximately half) of those obtained by SLEXbC. As we expected, for DbC and SLEXbC, error rates increase slightly with contamination A (weak) and notably

Figure 2.3: Boxplot of the misclassification rates in exercise 1, parameters values $+0.1$ versus 0



with contaminations B and C (strong), while changes are negligible for DbC- α because the trim keeps the contamination out. When contamination A is applied, DbC has about half the errors of SLEXbC, whereas their errors are similar with contaminations B and C. The three methods have no misclassifications for series that are easy to assign, that is, for values of ϕ far from 0. There are some symmetries in table 2.1 for DbC and SLEXbC: for example, the effect of contamination B with positive (negative) ϕ values of the autoregressive process model is similar to the effect of contamination C with negative (positive) ϕ . In addition, to extend the information provided by the tables, we include some boxplots showing distributions of the misclassification rates. For exercise 1, we only include the plot of one of the two most difficult discrimination settings, which is the comparison of the autoregressive model with $\phi = +0.1$ and the Gaussian white noise (see figure 2.3). The plot shows that SLEXbC tends to have a higher median, higher errors above this median, and fewer errors near zero. On the other hand, DbC- α is the only method that maintains the same pattern in the models with and without contamination and which has a considerable number of errors close to zero.

Tables 2.2, 2.3, 2.4 and 2.5 provide the results of the second simulation exercise. As expected, the errors decrease when any parameter, n or T , increases. The errors based on our methods, DbC and DbC- α , are larger than the errors using SLEXbC when we consider the whole series

Table 2.1: Misclassification rate estimates for simulation exercise 1 with and without contamination

	$\phi = -0.5$	$\phi = -0.3$	$\phi = -0.1$	$\phi = +0.1$	$\phi = +0.3$	$\phi = +0.5$
Without contamination						
DbC	0.000 (0.0000)	0.000 (0.0000)	0.063 (0.0017)	0.060 (0.0017)	0.000 (0.0000)	0.000 (0.0000)
DbC-α	0.000 (0.0000)	0.000 (0.0000)	0.065 (0.0018)	0.062 (0.0017)	0.000 (0.0000)	0.000 (0.0000)
SLEXbC	0.000 (0.0000)	0.000 (0.0000)	0.131 (0.0024)	0.127 (0.0024)	0.000 (0.0000)	0.000 (0.0000)
Contamination A						
DbC	0.000 (0.0000)	0.000 (0.0001)	0.077 (0.0019)	0.074 (0.0019)	0.000 (0.0001)	0.000 (0.0000)
DbC-α	0.000 (0.0000)	0.000 (0.0000)	0.064 (0.0017)	0.062 (0.0017)	0.000 (0.0000)	0.000 (0.0000)
SLEXbC	0.000 (0.0000)	0.000 (0.0001)	0.175 (0.0028)	0.172 (0.0029)	0.000 (0.0001)	0.000 (0.0000)
Contamination B						
DbC	0.000 (0.0000)	0.000 (0.0001)	0.300 (0.0028)	0.513 (0.0012)	0.001 (0.0002)	0.000 (0.0000)
DbC-α	0.000 (0.0000)	0.000 (0.0000)	0.065 (0.0018)	0.062 (0.0017)	0.000 (0.0000)	0.000 (0.0000)
SLEXbC	0.000 (0.0000)	0.001 (0.0002)	0.377 (0.0025)	0.491 (0.0011)	0.002 (0.0003)	0.000 (0.0000)
Contamination C						
DbC	0.000 (0.0000)	0.001 (0.0002)	0.512 (0.0013)	0.300 (0.0027)	0.000 (0.0001)	0.000 (0.0000)
DbC-α	0.000 (0.0000)	0.000 (0.0000)	0.064 (0.0017)	0.062 (0.0017)	0.000 (0.0000)	0.000 (0.0000)
SLEXbC	0.000 (0.0000)	0.002 (0.0004)	0.490 (0.0011)	0.377 (0.0025)	0.001 (0.0002)	0.000 (0.0000)

(without splitting them into blocks), although these errors fall with the first division. Notice that our methods outperform SLEXbc when series are divided into blocks, achieving the minimum error rates when $k = 2$. As we mentioned earlier, the length of the blocks decreases with k , and this implies that the quality of the estimated periodogram is decreased and the errors increase. This effect is reflected in all the tables and the optimal k is shown to be 2. Moreover, we can see that the increase in error with k is higher for short series than for longer ones. Recall that, like our procedure, the SLEXbC method implicitly splits the series into blocks. When we consider contaminations in the model, the error rates based on DbC and SLEXbC increase slightly with contamination A and greatly with contaminations B and C, while DbC- α maintains its errors and outperforms the other methods, especially with strong contaminations and $k = 2$. As expected, contaminating a series has major effects when samples sizes are $n = 8$ compared to when $n = 16$. The DbC and SLEXbC methods are more affected by contamination C than by contamination B, since $\phi = +0.9$ is farther from $\phi = -0.1$ (population $P^{(1)}$) than $\phi = -0.9$.

The boxplots of the error distributions for exercise 2 are represented in figure 2.4. As in the tables, the plots show that DbC and DbC- α perform better than SLEXbC when $k > 1$. The median error rate decreases when $k = 2$ (with respect to $k = 1$) and presents stable behaviour for k greater than two. These plots and tables reflect that DbC- α , with $k = 2$, tends to provide the best results, except when there is no contamination with which DbC with $k = 2$ outperforms all the other methods.

Similar results to the previous ones can be derived for simulation exercise 3 (see tables 2.6, 2.7, 2.8 and 2.9). They show that in our proposal the drawback of splitting too much is not relevant when series are long enough. With the presence of contamination, the best errors are obtained by DbC- α for $k = 4$. Contamination A has minor effects. On the other hand, results are very different for contaminations B and C. Notice that since τ has positive values in both populations, contaminating with a time series of parameter $\tau = -0.9$ (contamination B) has a stronger effect than using a series with $\tau = +0.9$ (contamination C).

Finally, in the three experiments a subtle effect can be seen between DbC and DbC- α . When there is no contamination it is normal for the former to provide slightly better error rates, because DbC- α is using only $100(1 - \alpha)\%$ of the training data available. Nevertheless, when there is some kind of contamination the best results are given by DbC- α .

Table 2.2: Misclassification rate estimates for simulation exercise 2 without contamination

	$n \times T = 8 \times 512$	16×512	8×1024	16×1024	8×2048	16×2048
DbC 1	0.141 (0.0024)	0.131 (0.0024)	0.062 (0.0017)	0.060 (0.0017)	0.014 (0.0008)	0.014 (0.0008)
2	0.066 (0.0017)	0.061 (0.0017)	0.015 (0.0009)	0.014 (0.0008)	0.001 (0.0003)	0.001 (0.0003)
4	0.078 (0.0019)	0.069 (0.0018)	0.015 (0.0009)	0.014 (0.0009)	0.001 (0.0003)	0.001 (0.0003)
8	0.090 (0.0020)	0.080 (0.0019)	0.020 (0.0010)	0.018 (0.0009)	0.002 (0.0003)	0.001 (0.0003)
DbC-α 1	0.143 (0.0024)	0.132 (0.0024)	0.063 (0.0017)	0.061 (0.0017)	0.015 (0.0009)	0.014 (0.0008)
2	0.069 (0.0018)	0.064 (0.0017)	0.016 (0.0009)	0.015 (0.0009)	0.001 (0.0003)	0.001 (0.0003)
4	0.083 (0.0020)	0.073 (0.0018)	0.017 (0.0010)	0.016 (0.0009)	0.002 (0.0003)	0.001 (0.0003)
8	0.105 (0.0023)	0.088 (0.0020)	0.024 (0.0011)	0.019 (0.0010)	0.002 (0.0004)	0.002 (0.0003)
SLEXbC	0.114 (0.0023)	0.086 (0.0020)	0.038 (0.0014)	0.025 (0.0011)	0.007 (0.0006)	0.003 (0.0004)

Table 2.3: Misclassification rate estimates for simulation exercise 2 with contamination A

	$n \times T = 8 \times 512$	16×512	8×1024	16×1024	8×2048	16×2048
DbC 1	0.143 (0.0025)	0.132 (0.0024)	0.063 (0.0017)	0.062 (0.0017)	0.018 (0.0010)	0.015 (0.0008)
2	0.070 (0.0018)	0.062 (0.0017)	0.018 (0.0010)	0.014 (0.0008)	0.002 (0.0003)	0.001 (0.0003)
4	0.083 (0.0020)	0.071 (0.0019)	0.019 (0.0010)	0.015 (0.0009)	0.002 (0.0003)	0.001 (0.0003)
8	0.102 (0.0022)	0.083 (0.0020)	0.026 (0.0012)	0.019 (0.0010)	0.003 (0.0004)	0.002 (0.0003)
DbC-α 1	0.145 (0.0025)	0.132 (0.0023)	0.063 (0.0017)	0.061 (0.0017)	0.015 (0.0009)	0.014 (0.0008)
2	0.072 (0.0018)	0.064 (0.0017)	0.015 (0.0009)	0.015 (0.0009)	0.001 (0.0002)	0.001 (0.0003)
4	0.086 (0.0021)	0.073 (0.0018)	0.018 (0.0010)	0.016 (0.0009)	0.002 (0.0003)	0.001 (0.0003)
8	0.114 (0.0024)	0.089 (0.0021)	0.025 (0.0011)	0.019 (0.0010)	0.003 (0.0004)	0.002 (0.0003)
SLEXbC	0.128 (0.0025)	0.092 (0.0021)	0.050 (0.0016)	0.027 (0.0012)	0.012 (0.0008)	0.004 (0.0004)

Table 2.4: Misclassification rate estimates for simulation exercise 2 with contamination B

	$n \times T = 8 \times 512$	16×512	8×1024	16×1024	8×2048	16×2048
DbC 1	0.258 (0.0029)	0.168 (0.0026)	0.252 (0.0029)	0.117 (0.0022)	0.250 (0.0029)	0.065 (0.0018)
2	0.135 (0.0024)	0.082 (0.0020)	0.088 (0.0021)	0.030 (0.0012)	0.049 (0.0016)	0.007 (0.0006)
4	0.137 (0.0025)	0.085 (0.0020)	0.089 (0.0021)	0.031 (0.0012)	0.049 (0.0016)	0.007 (0.0006)
8	0.143 (0.0025)	0.092 (0.0021)	0.093 (0.0022)	0.034 (0.0014)	0.050 (0.0016)	0.007 (0.0006)
DbC-α 1	0.145 (0.0024)	0.134 (0.0024)	0.064 (0.0017)	0.061 (0.0017)	0.015 (0.0008)	0.014 (0.0008)
2	0.070 (0.0018)	0.065 (0.0017)	0.017 (0.0010)	0.015 (0.0009)	0.003 (0.0006)	0.001 (0.0003)
4	0.081 (0.0020)	0.071 (0.0019)	0.017 (0.0010)	0.017 (0.0009)	0.002 (0.0003)	0.002 (0.0003)
8	0.104 (0.0023)	0.087 (0.0020)	0.023 (0.0011)	0.019 (0.0010)	0.002 (0.0004)	0.002 (0.0003)
SLEXbC	0.239 (0.0031)	0.134 (0.0024)	0.228 (0.0030)	0.081 (0.0020)	0.220 (0.0030)	0.037 (0.0013)

Table 2.5: Misclassification rate estimates for simulation exercise 2 with contamination C

	$n \times T = 8 \times 512$	16×512	8×1024	16×1024	8×2048	16×2048
DbC 1	0.457 (0.0056)	0.162 (0.0027)	0.437 (0.0055)	0.090 (0.0020)	0.445 (0.0047)	0.038 (0.0013)
2	0.147 (0.0036)	0.078 (0.0019)	0.055 (0.0020)	0.028 (0.0012)	0.015 (0.0010)	0.005 (0.0005)
4	0.187 (0.0037)	0.092 (0.0021)	0.068 (0.0022)	0.030 (0.0012)	0.017 (0.0010)	0.006 (0.0005)
8	0.225 (0.0039)	0.107 (0.0022)	0.101 (0.0027)	0.034 (0.0014)	0.024 (0.0011)	0.006 (0.0006)
DbC-α 1	0.145 (0.0025)	0.133 (0.0024)	0.063 (0.0017)	0.062 (0.0017)	0.015 (0.0009)	0.014 (0.0008)
2	0.073 (0.0020)	0.065 (0.0017)	0.018 (0.0013)	0.015 (0.0009)	0.002 (0.0005)	0.001 (0.0003)
4	0.083 (0.0020)	0.073 (0.0018)	0.017 (0.0010)	0.016 (0.0009)	0.002 (0.0003)	0.001 (0.0003)
8	0.108 (0.0022)	0.088 (0.0021)	0.024 (0.0011)	0.019 (0.0010)	0.003 (0.0004)	0.002 (0.0003)
SLEXbC	0.376 (0.0036)	0.177 (0.0029)	0.354 (0.0032)	0.098 (0.0023)	0.369 (0.0030)	0.040 (0.0015)

Figure 2.4: Boxplots of the misclassification error rates for simulation exercise 2, training sets with 8 series of length 1024

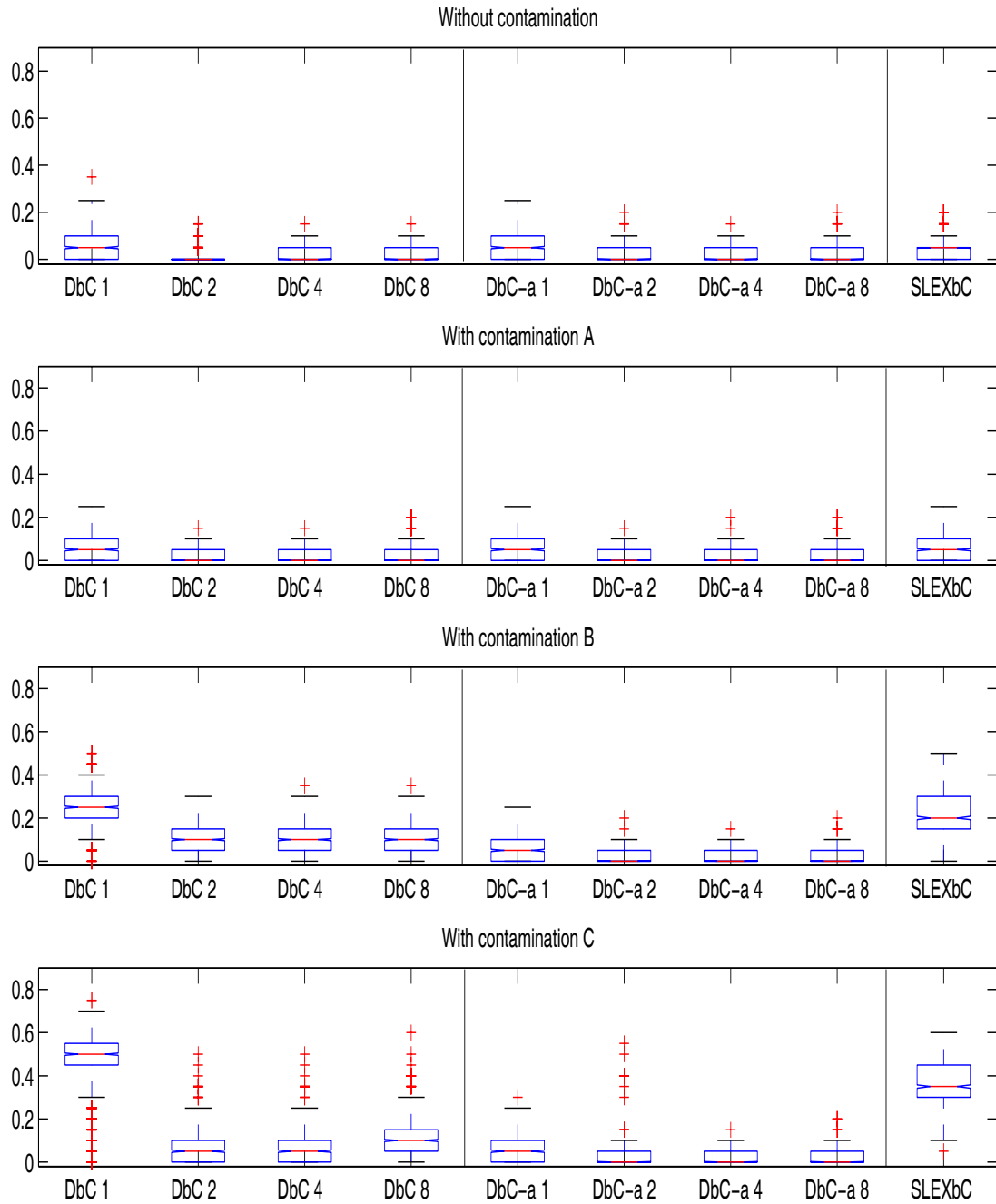


Table 2.6: Misclassification rate estimates for simulation exercise 3 without contamination

	$\tau = \mathbf{0.4}$	$\tau = \mathbf{0.3}$	$\tau = \mathbf{0.2}$
DbC 1	0.218 (0.0031)	0.063 (0.0017)	0.019 (0.0010)
2	0.119 (0.0023)	0.006 (0.0006)	0.000 (0.0000)
4	0.101 (0.0022)	0.002 (0.0003)	0.000 (0.0000)
8	0.123 (0.0024)	0.003 (0.0004)	0.000 (0.0000)
DbC-α 1	0.226 (0.0032)	0.065 (0.0018)	0.021 (0.0010)
2	0.128 (0.0023)	0.006 (0.0006)	0.000 (0.0000)
4	0.112 (0.0023)	0.002 (0.0003)	0.000 (0.0000)
8	0.139 (0.0026)	0.004 (0.0004)	0.000 (0.0000)
SLEXbC	0.181 (0.0031)	0.011 (0.0009)	0.000 (0.0000)

Table 2.7: Misclassification rate estimates for simulation exercise 3 with contamination A

	$\tau = \mathbf{0.4}$	$\tau = \mathbf{0.3}$	$\tau = \mathbf{0.2}$
DbC 1	0.232 (0.0032)	0.062 (0.0017)	0.019 (0.0009)
2	0.143 (0.0026)	0.006 (0.0006)	0.000 (0.0000)
4	0.144 (0.0026)	0.004 (0.0004)	0.000 (0.0000)
8	0.177 (0.0028)	0.005 (0.0005)	0.000 (0.0000)
DbC-α 1	0.241 (0.0035)	0.065 (0.0018)	0.020 (0.0010)
2	0.131 (0.0025)	0.007 (0.0006)	0.000 (0.0000)
4	0.121 (0.0026)	0.003 (0.0004)	0.000 (0.0000)
8	0.150 (0.0029)	0.005 (0.0005)	0.000 (0.0000)
SLEXbC	0.234 (0.0033)	0.016 (0.0011)	0.000 (0.0000)

Table 2.8: Misclassification rate estimates for simulation exercise 3 with contamination B

	$\tau = \mathbf{0.4}$	$\tau = \mathbf{0.3}$	$\tau = \mathbf{0.2}$
DbC 1	0.254 (0.0029)	0.106 (0.0022)	0.043 (0.0015)
2	0.500 (0.0015)	0.067 (0.0021)	0.001 (0.0002)
4	0.500 (0.0012)	0.062 (0.0020)	0.001 (0.0002)
8	0.499 (0.0013)	0.082 (0.0024)	0.000 (0.0001)
DbC-α 1	0.231 (0.0031)	0.074 (0.0020)	0.026 (0.0012)
2	0.128 (0.0024)	0.007 (0.0006)	0.000 (0.0000)
4	0.113 (0.0023)	0.002 (0.0004)	0.000 (0.0000)
8	0.141 (0.0026)	0.003 (0.0004)	0.000 (0.0000)
SLEXbC	0.492 (0.0019)	0.174 (0.0051)	0.015 (0.0009)

Table 2.9: Misclassification rate estimates for simulation exercise 3 with contamination C

	$\tau = \mathbf{0.4}$	$\tau = \mathbf{0.3}$	$\tau = \mathbf{0.2}$
DbC 1	0.257 (0.0029)	0.107 (0.0022)	0.044 (0.0015)
2	0.153 (0.0025)	0.017 (0.0009)	0.000 (0.0001)
4	0.128 (0.0024)	0.007 (0.0006)	0.000 (0.0000)
8	0.132 (0.0024)	0.006 (0.0006)	0.000 (0.0001)
DbC-α 1	0.234 (0.0031)	0.074 (0.0020)	0.025 (0.0012)
2	0.125 (0.0024)	0.007 (0.0006)	0.000 (0.0001)
4	0.114 (0.0024)	0.002 (0.0004)	0.000 (0.0000)
8	0.138 (0.0026)	0.004 (0.0004)	0.000 (0.0000)
SLEXbC	0.173 (0.0027)	0.015 (0.0009)	0.000 (0.0001)

Table 2.10: Mean computation times for simulation exercise 1

	$\phi = -0.5$	$\phi = -0.3$	$\phi = -0.1$	$\phi = +0.1$	$\phi = +0.3$	$\phi = +0.5$
DbC	0.027	0.027	0.027	0.027	0.027	0.027
DbC-α	0.044	0.045	0.045	0.044	0.044	0.044
SLEXbC	0.632	0.678	0.724	0.713	0.670	0.619

COMMENTS ON COMPUTATION TIMES

Estimates of the computation times are given in tables 2.10, 2.11 and 2.12. The computation time depends on the implementation—not just on the method itself—so we pay closer attention to the qualitative interpretation of the results, as they are less dependent on the programmed code.

Since the chronometer is called after generating the series, it can be expected that the computation times do not depend on the parameters of the stochastic processes. This is what we observed for our algorithms, but not for the SLEXbC method. Perhaps this is because this method needs to select a basis of the SLEX library for each series, while our method works only with the graphs of the functions and the computation of the integrated periodograms, which do not depend on the parameters.

Some conclusions that can be derived from the three simulation exercises are the following. It is clear that for our procedures, computation time increases with the number of blocks k . Also, table 2.11 shows that our methods, especially DbC- α , depend on sample size. The computation of depth is moderately time-consuming with the sample size and in less degree with series length. However, we have conveniently implemented the notion of depth in López-Pintado and Romo (2006) so it is computationally feasible and applicable to high sample sizes. Table 2.11 illustrates that DbC- α computation time increases with size but it is still reasonable and faster than SLEXbC. In short, for our approach, computation time depends more on the number of blocks, k , and the sample size, n , but not so much on the series length, T . In contrast, SLEXbC computation time depends on both, n and T , and increases when either of them increases.

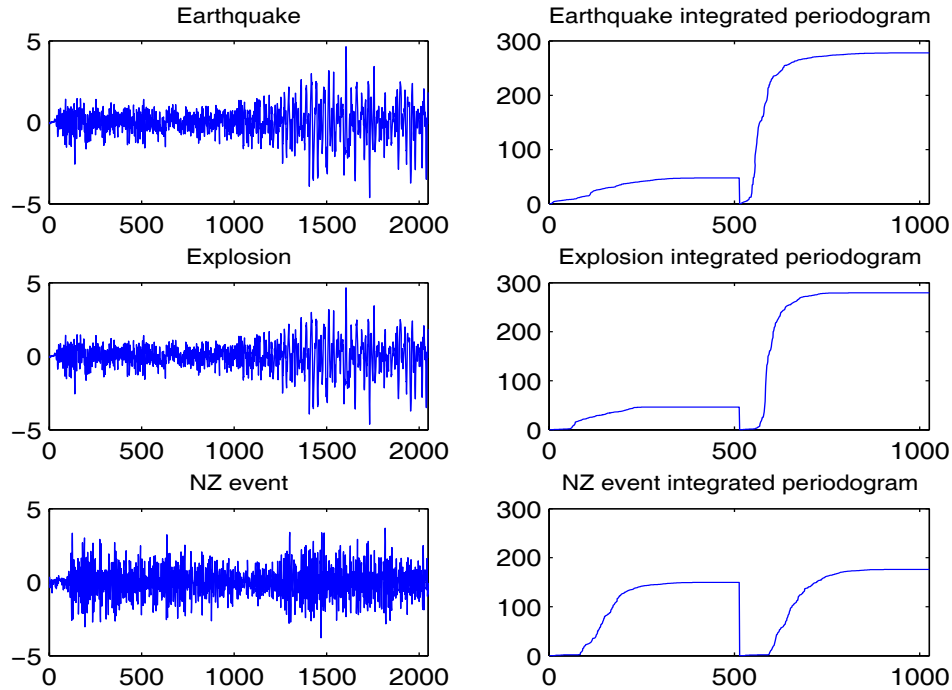
Table 2.11: Mean computation times for simulation exercise 2

	$n \times T = 8 \times 512$	16×512	8×1024	16×1024	8×2048	16×2048
DbC 1	0.021	0.028	0.027	0.038	0.044	0.067
2	0.036	0.049	0.043	0.060	0.062	0.087
4	0.066	0.092	0.067	0.094	0.081	0.115
8	0.125	0.180	0.126	0.181	0.129	0.186
DbC-α 1	0.031	0.108	0.044	0.200	0.084	0.463
2	0.046	0.137	0.064	0.237	0.103	0.496
4	0.086	0.280	0.087	0.276	0.123	0.505
8	0.170	0.585	0.171	0.595	0.173	0.602
SLEXbC	0.355	0.517	0.736	1.095	1.681	2.506

Table 2.12: Mean computation times for simulation exercise 3

	$\tau = 0.4$	$\tau = 0.3$	$\tau = 0.2$
DbC 1	0.031	0.030	0.030
2	0.047	0.047	0.048
4	0.074	0.074	0.074
8	0.140	0.140	0.140
DbC-α 1	0.066	0.062	0.063
2	0.083	0.093	0.094
4	0.120	0.121	0.120
8	0.235	0.234	0.235
SLEXbC	0.733	0.685	0.675

Figure 2.5: Real data examples and its curves



2.5 Real Data Example

2.5.1 Explosions and Earthquakes Data

We have evaluated our proposal in a benchmark data set containing eight explosions, eight earthquakes and one extra series —known as *NZ event*— not classified (but being either an earthquake or an explosion). This data set was constructed by Blandford (1993). Each series contains 2048 points, and its plot clearly shows two different parts — the first half is the part P and the second half is S. This division is an assumption considered by most authors, and it is based on geological reasons. It is also frequently considered that both parts are stationary. Kakizawa et al. (1998) give a list of these measurements. Shumway and Stoffer (2000) included a detailed study of this data set and provide access to the data set on the web site of their book: <http://www.stat.pitt.edu/stoffer/tsa.html>. Figure 2.5 presents examples of an earthquake and an explosion, and the NZ event.

Following the criterion to choose between normalized and nonnormalized versions of the cumulative periodogram given in section 2.2, we have considered the curve formed by merging the nonnormalized integrated periodograms of parts P and S independently computed; that is, we take $k = 2$. Let us consider the 8 earthquakes as group 1 and the eight explosions as group 2. We

have used leave-one-out cross validation to classify the elements of these two groups by removing a series at a time and using the rest of the data to train the method for final classification of the series. With this validation procedure, our two algorithms misclassify the first series of group 2 (explosions). Regarding the NZ event, both algorithms agree on assigning it to the explosions group, as described previously by other authors (for example, Kakizawa et al. [1998], Huang et al. [2004]).

An additional exercise considers an artificial data set constructed by the eight earthquakes plus the NZ event as group 1, and the eight explosions as group 2. Note that our method and most of the published papers classify NZ as an explosion. Therefore, this could be considered an artificial scenario where an outlier is presented in group 1. In this situation, algorithm 1 misclassifies the first and the third elements of group 2 (explosions), whereas algorithm 2 misclassifies only the first series of group 2. This seems to show the robustness of our second algorithm. Obviously, as we are using leave-one-out cross validation, both algorithms classify the NZ event in the explosions group.

2.6 Conclusions

We propose a new frequency domain approach for time series classification based on the integrated periodograms of the series. When series are nonstationary, they are split into blocks and the integrated periodograms of the blocks are merged to construct a curve. This idea relays on the assumption that series are locally stationary. Since the integrated periodogram is a function, the statistical tools for functional data analysis can be applied. In our classification procedure new series are assigned to the class minimizing the distance between its corresponding curve and the group mean curve. Since the group mean can be affected by the presence of outliers, robustness of the classification method is achieved by substituting the mean curve with the α -trimmed mean, where for each group only the deepest elements are averaged. To evaluate our proposal in different scenarios, we have done simulation exercises containing several models and parameters, with both stationary and nonstationary series, as well as with different types of contamination. We have also illustrated the performance of our procedure in a real benchmark data set. Our proposal provides small error rates, robustness, and good computational behaviour, properties which make the methodology suitable for time series classification. It also outperforms previous methods

proposed in the literature. This chapter suggests that the integrated periodogram contains useful information for classifying time series.

Chapter 3

Functional Data Classification

Summary: A popular approach for classifying functional data is based on the distances from the function or its derivatives to group representative (usually the mean) functions or their derivatives. In this chapter, we propose using a combination of those distances. Simulation studies show that our procedure performs very well, resulting in smaller testing classification errors. Applications to real data show that our procedure performs as well as —and in some cases better than— other classification methods.

Key words: discriminant analysis, functional data, weighted distances.

3.1 Introduction

Functional data have great —and growing— importance in Statistics. Nowadays, functional data are present in many areas, sometimes because they are the output of measurement processes, other times for theoretical or practical reasons; functional models are used even for nonfunctional data (see section 1.2 of Ramsay and Silverman [2006]). Most of the classical techniques for the finite- and high-dimensional frameworks have been adapted to cope with the infinite dimensions, but due to the *curse of dimensionality*, new and specific treatments are still required. As with other types of data, statisticians must supervise different steps —registration, missing data, representation, transformation, typicality— and tackle different tasks —modelization, classification or clustering, among others. In practice, curves can neither be registered continuously nor at infinite points. Then, techniques dealing with high-dimensional data can sometimes be applied: Hastie

et al. (1995), for example, adapt the discriminant analysis to cope with many highly correlated predictors, “such as those obtained by discretizing a function”.

Among the approaches specifically designed for functional data classification, the following project the data into a finite-dimensional space of functions and therefore work with the coefficients; this technique is called *filtering*. James and Hastie (2001) model the coefficients with “Gaussian distribution with common covariance matrix for all classes, by analogy with LDA [linear discriminant analysis]”; their classification minimizes the distance to the group mean. The classification method of Hall et al. (2001) maximizes the likelihood, and although they propose a fully nonparametric density estimation, in practice multivariate Gaussian densities are considered, leading to quadratic discriminant analysis. Biau et al. (2003) apply k -nearest neighbour to the coefficients, while Rossi and Villa (2006) apply support vector machines. Berlinet et al. (2008) extend the approach of Biau et al. (2003) to wavelet bases and to more general discrimination rules. The following proposals are designed to make direct use of the continuity of the functional data. Ferraty and Vieu (2003) classify new curves in the group with the highest posterior probability of membership kernel estimate. On the other hand, López-Pintado and Romo (2006) also take into account the continuity feature of the data and propose two classification methods based on the notion of *depth* for curves; in their first proposal new curves are assigned to the group with the closest trimmed mean, while the second method minimizes a weighted average distance to each element in the group. Abraham et al. (2006) extend the moving window rule for functional data classification. Nerini and Ghattas (2007) classify density functions with functional regression trees. Baíllo and Cuevas (2008) provide some theoretical results on the functional k -nearest neighbour classifier, and suggest —as a partial answer— that this method could play the same central role for functional data as Fisher’s method for the finite-dimensional case. To use only the most informative parts of the curves, Li and Yu (2008) have proposed a new idea: they use F-statistics to select the place where linear discriminant analysis is applied into small intervals, providing an output that is used as input in a final support vector machines step.

There are several works addressing the unsupervised classification —or clustering— problem. Abraham et al. (2003) fit the functional data by B-splines and apply k -means on the coefficients. James and Sugar (2003) project the data into a finite-dimensional space and consider a random-effects model for the coefficients; their method is effective when the observations are sparse, irregularly spaced or occur at different time points for each subject. The continuous nature of the

data is used, in a more direct form, by the following works. The proposal of Tarpey and Kinader (2003) classifies using a k -means algorithm over the probability distributions. A hierarchical descending procedure, using heterogeneity indexes based on modal and mean curves, is presented in Dabo-Niang et al. (2006). Impartial trimming is combined with k -means in Cuesta-Albertos and Fraiman (2007).

Functional data can be transformed in several ways. After the registration, spatial or temporal alignments are sometimes necessary; references on this topic are Wang and Gasser (1997, 1999) and Ramsay and Silverman (2006). On the other hand, Dabo-Niang et al. (2007) use a distance invariant to small shifts. Examples of centering, normalization and derivative transformations are found in Rossi and Villa (2006). The objective of the transformations is to highlight some features of the data and to allow the information to be used more efficiently. For this kind of data, the most important transformation is taking derivatives. Since the different derivatives can contribute new information, a possible combination of them —or their information— should be taken into account. Mathematical *Functional Analysis* has been working with such combinations for a long time, mainly through some norms (in norm and Sobolev spaces), and Ramsay and Silverman (2006) find them frequently as a consequence of model adjustments or system properties (for Canadian weather stations data, melanoma data or lower lip movement data).

In order to obtain semimetrics, instead of metrics, Ferraty and Vieu (2006) consider derivatives (one at a time) in the distances. This implies theoretical advantages —throughout the topological structure induced by the semimetric— in the small ball probability function, providing a new way to deal with the curse of dimensionality.

We transform the functional data classification problem into a classical multivariate data classification problem. While the filtering techniques encapsulate the functional information into a set of coefficients, we construct a linear combination of variables and coefficients. Given the variables, the *linear discriminant analysis* determines the combination. Our proposal is based on the interpretation as variables of the distances between a new curve and the transformed and untransformed functional data. On the one hand, the classification can be improved, and, on the other hand, the coefficients of the combination provide information about the importance of each data transformation. When a nonnegativeness condition is applied to the coefficients, the combination (discriminant function) can be interpreted as the difference of measurements with a weighted distance. This metric automatically becomes a semimetric when the importance of the

distance to the untransformed data is null or insignificant; but the user can force, by considering only the derivatives as input, the method to output a semimetric.

The chapter is organized as follows. In section 2 the classification method is presented and described, from the optimization problem to the classification algorithm. In section 3, our proposal is evaluated with several simulation exercises. Two real data sets are classified in section 4. Finally, in section 5 a summary of conclusions is given.

3.2 The Classification Method

3.2.1 The Optimization Problem

AN ADDITIONAL CONSTRAINT

In order to base the classification on a semimetric or on a metric, one version of our proposal adds another constraint—in fact, several nonnegativity constraints—to the classical Fisher's discriminant analysis optimization problem:

$$\mathbf{a} = \operatorname{argmax} \{ \mathbf{a}^t \mathbf{B} \mathbf{a} \} \quad \text{subject to} \quad \begin{cases} \mathbf{a}^t \mathbf{W} \mathbf{a} = 1 \\ \mathbf{a} \geq \mathbf{0} \end{cases}, \quad (3.1)$$

where \mathbf{B} is the *between-class scatter matrix*, \mathbf{W} is the *within-class scatter matrix*, $\mathbf{a} = (a_1, \dots, a_p)^t$, and $\mathbf{a} \geq \mathbf{0}$ means $a_i \geq 0$, $i = 1, \dots, p$ (see appendix B for the definition of \mathbf{B} and \mathbf{W} . Remind that \mathbf{B} is positive semidefinite and \mathbf{W} is, by hypothesis, positive definite). This is a nonlinear (quadratic) programming problem with an *equality constraint* and *nonnegativity constraints*. The latter constraints are frequently dealt with in literature, since they appear naturally when considering the dual problems of linear and quadratic programs (see examples 3.4.2 and 3.4.3 of Bertsekas [1999] or sections 4.3 and 4.4 of Boyd and Vandenberghe [2008]). The solution of this new optimization problem can be represented by the pair $(\mathbf{a}_p, \lambda_p)$, with $\mathbf{a}_p^t \mathbf{W} \mathbf{a}_p = 1$, $\mathbf{a}_p \geq \mathbf{0}$ and $\lambda_p = \mathbf{a}_p^t \mathbf{B} \mathbf{a}_p$. Let us denote $V_{\mathbf{a}}^* = \{c\mathbf{a}, \ c \in \mathbb{R}, \ c \neq 0\}$. Section 3.2.2 contains some theory on obtaining the explicit expression of \mathbf{a}_p .

Geometrically, the set $V_{\mathbf{a}} = V_{\mathbf{a}}^* \cup \{\mathbf{a} = \mathbf{0}\}$ is a one-dimensional linear subspace of \mathbb{R}^p . When $V_{\mathbf{a}}^*$ intersects the nonnegative orthant $\{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a} \geq \mathbf{0}\}$ outside the origin, this last optimization problem will provide the same solution as those without the nonnegativity constraints.

CONVEXITY

The cost function $\mathbf{a}^t \mathbf{B} \mathbf{a}$ is convex due to proposition 13.d and the positive definiteness of the matrix \mathbf{B} .

EXISTENCE OF SOLUTIONS

In this new optimization problem, the feasible domain is

$$\mathcal{D} = \{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a}^t \mathbf{W} \mathbf{a} = 1 \text{ and } \mathbf{a} \geq \mathbf{0}\}, \quad (3.2)$$

that is convex, since proposition 14.a can be applied after writing

$$\mathcal{D} = \{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\} \cap \{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a} \geq \mathbf{0}\}. \quad (3.3)$$

On the other hand, as the inequality constraints are expressed in terms of linear functions, they do not change the behaviour of the second derivatives of the Lagrangian, that is, do not change the convexity of the cost function of the optimization problem. As a consequence, the existence of solution is guaranteed as in the classical discriminant analysis optimization problem (see appendix B).

CASE $K = 2$: TWO POPULATIONS

In this case, the optimization problem (3.1) is equivalent to the following one (see appendix B):

$$\mathbf{a} = \operatorname{argmax} \{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2\} \quad \text{subject to} \quad \begin{cases} \mathbf{a}^t \mathbf{W} \mathbf{a} = 1 \\ \mathbf{a} \geq \mathbf{0} \end{cases}. \quad (3.4)$$

3.2.2 The Discriminant Function

The expression of the discriminant function with our additional constraint, $y = \mathbf{a}_p^t \mathbf{x}$, is more difficult to obtain than in the classical case (see appendix B). We present explicit expressions for some specific easy cases (notice that in this work we consider the cases $p = 1, 2$ or 3). Although we are interested in the $K = 2$ case, some of the following calculations are made with the same difficulty for the general K -populations case: that is, for the general problem (3.1) instead of this particular one (3.4). As was mentioned, when the linear subspace E_λ^* of \mathbb{R}^p intersects the nonnegative orthant $\{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a} \geq \mathbf{0}\}$ outside the origin, that is, when all the components of \mathbf{a}_F

have the same sign, the new discriminant function will be

$$y = \mathbf{a}_p^t \mathbf{x} = \alpha \mathbf{a}_F^t \mathbf{x} = \alpha (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \mathbf{x}, \quad (3.5)$$

with $\alpha = +1$ or $\alpha = -1$ so that the condition $\alpha \mathbf{a}_F \geq \mathbf{0}$ holds.

In general, when all the components of \mathbf{a}_F do not have the same sign, formal calculations are necessary. The *objective function* —of the optimization problem— and the constraints are combined in the *Lagrangian*, and the nonnegativeness is taken into account through the *Karush-Kuhn-Tucker conditions*, that are necessary and sufficient (see proposition 26):

$$\begin{cases} \frac{\partial L}{\partial \mathbf{a}} = \mathbf{0} \\ \frac{\partial L}{\partial \beta} = \mathbf{0} \\ a_i \geq 0, \mu_i \geq 0 \text{ and } \mu_i a_i = 0, \end{cases} \quad (3.6)$$

where the Lagrangian is

$$L(\mathbf{a}, \beta, \mu) = \mathbf{a}^t \mathbf{B} \mathbf{a} + \beta(1 - \mathbf{a}^t \mathbf{W} \mathbf{a}) + \mathbf{a}^t \mu \quad (3.7)$$

and $\mu = (\mu_1, \dots, \mu_p)^t$ and β are the *multipliers*. It holds that

$$\frac{\partial L}{\partial \mathbf{a}} = 2\mathbf{B} \mathbf{a} - \beta 2\mathbf{W} \mathbf{a} + \mu. \quad (3.8)$$

The conditions (3.6) become

$$\begin{cases} 2(\mathbf{B} - \beta \mathbf{W}) \mathbf{a} = -\mu \\ \mathbf{a}^t \mathbf{W} \mathbf{a} = 1 \\ a_i \geq 0, \mu_i \geq 0 \text{ and } \mu_i a_i = 0, \end{cases} \quad (3.9)$$

that are a system with $2p + 1$ conditions and variables. Giving explicit solution of this system is only possible in some simple cases.

CASE $p = 1$: ONE VARIABLE

In this case, with only one discriminant variable, the original Fisher's discriminant analysis is trivial, since

$$\lambda(a) = \frac{aBa}{aWa} = \frac{B}{W} = \text{constant}. \quad (3.10)$$

CASE $p = 2$: TWO VARIABLES

First of all, when **two populations** are considered, let us denote

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix},$$

where by definition $w_{12} = w_{21}$ and $b_{12} = b_{21}$.

For two discriminant variables, three nonnull subcases (see the searching strategy mentioned in appendix C) cover the bidimensional positive quadrant where \mathbf{a} is:

(A) Case $a_1 > 0$, $\mu_1 = 0$ and $a_2 = 0$. In this case,

(a1) By hypothesis, $\mu_1 = 0$ and $a_2 = 0$.

(a2) From $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$ the value $a_1 = |\sqrt{w_{11}^{-1}}|$ is obtained.

(a3) Finally, $2(\mathbf{B} - \beta \mathbf{W}) \mathbf{a} = -\mu$ implies that $\beta = w_{11}^{-1} b_{11}$ and $\mu_2 = -2(b_{21} - w_{11}^{-1} b_{11} w_{21}) |\sqrt{w_{11}^{-1}}|$.

The discriminant function, if $\mathbf{B} - \beta \mathbf{W}$ is negative semidefinite, would be

$$y_A = \mathbf{a}_p^t \mathbf{x} = |\sqrt{w_{11}^{-1}}| x_1. \quad (3.11)$$

(B) Case $a_1 = 0$, $a_2 > 0$ and $\mu_2 = 0$. In this case,

(b1) By hypothesis, $a_1 = 0$ and $\mu_2 = 0$.

(b2) Now, $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$ implies the value $a_2 = |\sqrt{w_{22}^{-1}}|$.

(b3) From $2(\mathbf{B} - \beta \mathbf{W}) \mathbf{a} = -\mu$ the values $\beta = w_{22}^{-1} b_{22}$ and $\mu_1 = -2(b_{12} - w_{22}^{-1} b_{22} w_{12}) |\sqrt{w_{22}^{-1}}|$ are obtained.

The discriminant function, if $\mathbf{B} - \beta \mathbf{W}$ is negative semidefinite, would be

$$y_B = \mathbf{a}_p^t \mathbf{x} = |\sqrt{w_{22}^{-1}}| x_2. \quad (3.12)$$

(C) Case $a_1 > 0$, $\mu_1 = 0$, $a_2 > 0$ and $\mu_2 = 0$. To study the *interior solutions*,

(c1) By hypothesis $\mu = \mathbf{0}$, the nonnegativity constraint disappears from the Lagrangian and the objective function is again $L(\mathbf{a}) = \lambda(\mathbf{a})$.

(c2) As $(\mathbf{B} - \beta\mathbf{W})\mathbf{a} = \mathbf{0}$, it is necessary that $|\mathbf{B} - \beta\mathbf{W}| = 0$; this condition implies, since \mathbf{W} is not singular (by hypothesis), that

$$\beta = \frac{-b \pm \sqrt{b^2 - 4|\mathbf{W}||\mathbf{B}|}}{2|\mathbf{W}|}, \quad (3.13)$$

with $b = w_{12}b_{21} + w_{21}b_{12} - w_{11}b_{22} - w_{22}b_{11}$. This means that $(\mathbf{W}^{-1}\mathbf{B} - \beta\mathbf{I})\mathbf{a} = \mathbf{0}$ and we are again interested in an eigenvector of an eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. Nevertheless, now the criterion is not selecting the largest eigenvalue, but selecting the largest one with eigenvectors verifying the nonnegativity constraint (or nonpositiveness, since the scale factor is not a problem).

(c3) Given β , also from $(\mathbf{B} - \beta\mathbf{W})\mathbf{a} = \mathbf{0}$ there will be nontrivial solution if $a_2 = \gamma a_1$, or, equivalently, $a_1 = \gamma^{-1}a_2$ with

$$\gamma = -\frac{b_{11} - \beta w_{11}}{b_{12} - \beta w_{12}}, \quad \text{or, equivalently,} \quad \gamma = -\frac{b_{21} - \beta w_{21}}{b_{22} - \beta w_{22}}, \quad (3.14)$$

as $|\mathbf{B} - \beta\mathbf{W}| = 0$.

(c4) Finally, the condition $\mathbf{a}^t\mathbf{W}\mathbf{a} = 1$ implies that

$$a_1 = |\sqrt{[w_{11} + \gamma(w_{12} + w_{21}) + \gamma^2 w_{22}]^{-1}}|, \quad (3.15)$$

or, respectively,

$$a_2 = |\sqrt{[\gamma^{-2}w_{11} + \gamma^{-1}(w_{12} + w_{21}) + w_{22}]^{-1}}|, \quad (3.16)$$

so the discriminant function, if $\mathbf{B} - \beta\mathbf{W}$ is negative semidefinite, would be

$$y_C = \mathbf{a}_p^t \mathbf{x} = a_1 x_1 + \gamma a_1 x_2, \quad (3.17)$$

or, respectively,

$$y_C = \mathbf{a}_p^t \mathbf{x} = \gamma^{-1} a_2 x_1 + a_2 x_2, \quad (3.18)$$

with γ (and β) as given above.

Remark 17 In the last expressions it has been implicitly supposed that $\gamma \neq 0$ and $\gamma \neq \infty$. It is, however, noteworthy that when $\gamma \rightarrow 0$ or $\gamma \rightarrow \infty$, the discriminant functions of the cases (A) and (B) arise, respectively, as limit cases of (C). As $a_1 \rightarrow |\sqrt{w_{11}^{-1}}|$ when $\gamma \rightarrow 0$ and $a_2 \rightarrow |\sqrt{w_{22}^{-1}}|$ when $\gamma \rightarrow \infty$, respectively, then

$$y_C \xrightarrow[\gamma \rightarrow 0]{} y_A \quad \text{and} \quad y_C \xrightarrow[\gamma \rightarrow \infty]{} y_B. \quad (3.19)$$

Remark 18 The parameter γ acquires an important role, since it provides information —under the nonnegativity constraints— about each variable importance for classifying purposes, that is, about each variable discriminant power.

Remark 19 This simple case, $p = 2$ (two variables), can be used to understand better the meaning of the within-class scatter matrix. By definition,

$$\mathbf{W} = \sum_{k=1}^K n_k \hat{\Sigma}_{\mathbf{x}}^{(k)} = \sum_{k=1}^K n_k (\hat{\sigma}_{ij}^{(k)})_{i,j} = \left(\sum_{k=1}^K n_k \hat{\sigma}_{ij}^{(k)} \right)_{i,j}, \quad (3.20)$$

where $\hat{\sigma}_{ij}^{(k)} = n_k^{-1} \sum_{e=1}^{n_k} (x_{ie}^{(k)} - \bar{x}_i^{(k)})(x_{je}^{(k)} - \bar{x}_j^{(k)})$. Then, for K populations,

$$\mathbf{W} = (w_{ij})_{i,j} = \left(\sum_{k=1}^K \sum_{e=1}^{n_k} (x_{ie}^{(k)} - \bar{x}_i^{(k)})(x_{je}^{(k)} - \bar{x}_j^{(k)}) \right)_{i,j}, \quad (3.21)$$

and, for two populations and two variables,

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = \begin{pmatrix} n_1 \hat{\sigma}_{11}^{(1)} + n_2 \hat{\sigma}_{11}^{(2)} & n_1 \hat{\sigma}_{12}^{(1)} + n_2 \hat{\sigma}_{12}^{(2)} \\ n_1 \hat{\sigma}_{21}^{(1)} + n_2 \hat{\sigma}_{21}^{(2)} & n_1 \hat{\sigma}_{22}^{(1)} + n_2 \hat{\sigma}_{22}^{(2)} \end{pmatrix}. \quad (3.22)$$

Remark 20 The Karush-Kuhn-Tucker conditions are sufficient due to proposition 26; this means that the previous computes have led to the local and global minimum. On the other hand, as these conditions are also necessary, due to proposition 24, it holds that

$$\nabla_{\mathbf{aa}}^2 L(\mathbf{a}, \beta, \mu) = \frac{\partial^2 L}{\partial \mathbf{a}^2}(\mathbf{a}, \beta, \mu) = 2(\mathbf{B} - \beta \mathbf{W}) \quad (3.23)$$

is a positive semidefinite matrix (in this case, $p = 2$, and under the nonnegativity constraints).

OTHER VALUES OF p

From the Karush-Kuhn-Tucker conditions of the cases $p = 3$ or $p = 4$, several subcases would arise after some work, providing explicit expressions for \mathbf{a}_p under some conditions on the samples. Nevertheless, since it has been proved that there are no formula for the solution of a five-degree general polynomial equation, for the cases $p \geq 5$ it would be impossible to find —in this way— the explicit expressions for \mathbf{a}_p .

3.2.3 The Classification

To classify new elements, the previous discriminant function is applied following the same ideas than in the classical discriminant analysis (see section B.4). Geometrically, the condition $\mathbf{a} \geq 0$

restricts the possible directions into which the data should be projected. We also determine the cutoff point by projecting $\frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ with $y = \mathbf{a}_p^t \mathbf{x}$, that is, via the $\mathbf{a}_p^t \cdot$ premultiplication. The method classifies a new element in the population k as follows:

$$\begin{cases} k = 1 & \text{if } y > \frac{1}{2}\mathbf{a}_p^t(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\ k = 2 & \text{otherwise} \end{cases}, \quad (3.24)$$

where the value $\frac{1}{2}\mathbf{a}_p^t(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ can be termed the *adjusted cutoff point*. Notice that for the particular case (3.5) the classification is just the same as that of the classical discriminant analysis.

As for $y = \mathbf{a}_F^t \mathbf{x}$, now the classification of a multivariate point is done for $y = \mathbf{a}_p^t \mathbf{x}$ by the simple comparison of its projection with the projection of the semisum of the group means. The calculations with simulated and real data show that the classification provided by the two discriminant functions is similar, while the nonnegativity restriction adds some theoretical advantages.

Remark 21 For each k , the centroid $\bar{\mathbf{x}}^{(k)}$ of each group can be interpreted as a representative multivariate point of the group, $\mathbf{r}^{(k)} = \bar{\mathbf{x}}^{(k)}$. Then, the previous classification criterion can be expressed as

$$k = \operatorname{argmin}_{\{1,2\}} \{d(y(\mathbf{x}), y(\mathbf{r}^{(k)}))\}. \quad (3.25)$$

Equivalently, for the group k the quantity $y(\mathbf{r}^{(k)})$ can be thought of as a representative univariate point.

3.2.4 Our Discriminant Variables

In order to facilitate understanding of the classification criterion, so far we have used generic discriminant variables x_1, \dots, x_p . Now we define the specific variables and explain how to construct them from the functional data.

If $\chi^{(1)}$ and $\chi^{(2)}$ are $(p-1)$ -order differentiable functions in a functional space L , the quantities $d(D^i \chi^{(1)}, D^i \chi^{(2)})$, for $i = 0, 1, \dots, p-1$, are numeric when $d(\cdot, \cdot)$ is a distance and the D^i denotes the i -th derivative ($i = 0$ represents no differentiation).

Assuming that there are two populations, with models $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$, and let $\chi_1^{(k)}, \dots, \chi_{n_k}^{(k)}$ be a sample of the population k , respectively; in this situation, for a function χ we define the variables

$$x_i = d(D^{i-1} \chi, \overline{D^{i-1} \chi}^{(1)}) - d(D^{i-1} \chi, \overline{D^{i-1} \chi}^{(2)}), \quad (3.26)$$

for $i = 1, 2, \dots, p$, where $\overline{D^{i-1}\chi}^{(k)} = n_k^{-1} \sum_{e=1}^{n_k} D^{i-1}\chi_e^{(k)} = D^{i-1}(n_k^{-1} \sum_{e=1}^{n_k} \chi_e^{(k)})$, $k = 1, 2$. That is, x_i is the difference between the distances from $D^{i-1}\chi$ to the $(i-1)$ -th derivative of the population means. With these definitions, the discriminant analysis will provide information about the usefulness of each derivative for classification purposes.

Remark 22 In this chapter we do not highlight the population version of concepts, but, in terms of the models $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ our discriminant variables would be, for a stochastic function \mathcal{X} ,

$$X_i = d(D^{i-1}\mathcal{X}, \mathbb{E}(D^{i-1}\mathcal{X}^{(1)})) - d(D^{i-1}\mathcal{X}, \mathbb{E}(D^{i-1}\mathcal{X}^{(2)})). \quad (3.27)$$

STANDARDIZATION AND COEFFICIENTS

At this point, it is advisable to study the relationship between the variables just defined and the interpretation of the coefficients provided by the general optimization problem (see section B.2.3).

Supposing that a variable t and a function $\chi(t)$ are not dimensionless (scalars without units of measure), nor is $D^1\chi(t) = d\chi(t)/dt$. Besides, the derivative has a different dimension than its original function, as the term $d\chi(t)$ has the same units than $\chi(t)$ and the term dt does not. As a consequence, all the variables defined in (3.26) are dimensionless only when t and $\chi(t)$ also are.

Anyway, for classification and descriptive purposes the transformation and the standardization of the data must be applied, respectively, as explained in section B.2.3. In our methodology this could be done over the functions (definitions of mean and standard deviation for functional data are given in literature), but it is preferable to operate over the multivariate data, as they are just in the input of the multivariate optimization problem and it is not sure that the changes were preserved in the functional-to-multivariate data transformation step.

3.2.5 The Algorithm

ALGORITHM 3

Let $\chi_1^{(k)}(t), \dots, \chi_{n_k}^{(k)}(t)$, $k = 1, 2$, be samples of functions from the two populations, then:

1. **From functional to multivariate data.** For each $\chi_e^{(k)}(t)$, $e = 1, \dots, n_k$, the

following vector is constructed

$$\mathbf{x}_e^{(k)} = (x_{1,e}^{(k)}, \dots, x_{p,e}^{(k)})^t, \quad (3.28)$$

where $x_{i,e}^{(k)}$ is obtained by (3.26). These vectors form the multivariate sample

$$(\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_{n_k}^{(k)}). \quad (3.29)$$

2. **The discriminant function.** These samples are used as input in the optimization problem to obtain the discriminant function:

$$y(\mathbf{x}) = \mathbf{a}^t \mathbf{x}, \quad (3.30)$$

where $\mathbf{x} = (x_1, \dots, x_p)^t$, and $\mathbf{a} = \mathbf{a}_F$ or $\mathbf{a} = \mathbf{a}_p$ depending on whether or not the additional constraint was imposed.

3. **The allocation of new curves.** To classify a new curve $\chi(t)$, its multivariate vector is constructed,

$$\mathbf{x} = (x_1, \dots, x_p)^t, \quad (3.31)$$

again using expression (3.26), and finally the value $y(\mathbf{x})$ is used to assign the curve $\chi(t)$ to one of the two populations, as mentioned in subsections 3.2.3 and B.4.

Remark 23 As a distance measurement between two functions we have taken the distance given in expression (1.65) for $m = 1$.

Remark 24 The possible outliers in the samples of functions could be extracted with the same methodology used in section 2.3.

Several versions of this algorithm have been implemented and compared in the following sections.

3.2.6 Weighted Semidistances or Distances

Let us substitute, for a function $\chi(t)$, the discriminant variables into the expression of the discriminant function:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{a}^t \mathbf{x} = \sum_{i=1}^p a_i x_i \\ &= \sum_{i=1}^p a_i d(D^{i-1} \chi, \overline{D^{i-1} \chi}^{(1)}) - \sum_{i=1}^p a_i d(D^{i-1} \chi, \overline{D^{i-1} \chi}^{(2)}). \end{aligned} \quad (3.32)$$

For the linear combinations $\sum_{i=1}^p a_i d(D^{i-1}\chi, \overline{D^{i-1}\chi}^{(k)})$ to take nonnegative values, our additional restrictions ($a_i \geq 0$) are necessary; so only the function $y = \mathbf{a}_p^t \mathbf{x}$ —not the classical linear discriminant function— can be seen as providing a classification based on the minimization of a weighted distance.

As in a space of functions the derivation can imply a loss of information, then

$$\zeta(\chi, \overline{\chi}^{(k)}) = \sum_{i=1}^p a_i d(D^{i-1}\chi, \overline{D^{i-1}\chi}^{(k)}), \quad (3.33)$$

with $a_i \geq 0$ can be interpreted as measurements with a weighted distance if and only if $a_1 \neq 0$ (in practice, if and only if a_1 is significant); otherwise, it can be interpreted as a measurement with a weighted semidistance, since two functions can differ in a constant and verify that $\zeta(\chi^{(1)}, \chi^{(2)}) = 0$. An important general property of $\zeta(\cdot, \cdot)$ is that it takes into account at the same time the functions, their smoothness, their curvature, etcetera.

Similarly, when the distance $d(\cdot, \cdot)$ is defined from a norm, the expression (3.33) can be seen as a weighted norm if and only if $a_1 \neq 0$ (in practice, if and only if a_1 is significant), and as a weighted seminorm otherwise.

3.3 Simulation Results

In order to illustrate the behaviour of our two procedures, we perform a Monte Carlo study using three different settings. In all cases we consider two functional populations in the space $\mathcal{C}[0, 1]$ of continuous functions defined in the interval $[0, 1]$. The methods used to classify are the following:

- Distance to the sample functional mean calculated using the functions in the training set (*DFM0*). That is, using the rule

$$\begin{cases} k = 1 & \text{if } x_1 < 0 \\ k = 2 & \text{otherwise} \end{cases}. \quad (3.34)$$

- Distance to the sample functional mean calculated using the first derivatives of functions in the training set (*DFM1*). That is, using the rule

$$\begin{cases} k = 1 & \text{if } x_2 < 0 \\ k = 2 & \text{otherwise} \end{cases}. \quad (3.35)$$

- Weighted indicator (*WI*) obtained using our first procedure. Using the algorithm with $\mathbf{x} = (x_1, x_2)^t$ and without our additional constraint.
- Weighted distance (*WD*) obtained using our second procedure. Then, for the algorithm with $\mathbf{x} = (x_1, x_2)^t$ and the nonnegativity constraint.

We generate 200 functions from each population. The training set consists of the first 100 functions from each population, and the remaining 100 observations from each sample are the test set. For each setting we run 1000 replications, so the results are based on 1000 estimates of the misclassification rates.

Now, we describe the three considered settings.

Simulation Exercise 1. We consider the following two functional data generating models:

Model B1. $\mathcal{X}_e^{(1)} = t + U_e$, where U_e is a uniform random variable on the interval $(0, 1)$.

Model R1. $\mathcal{X}_e^{(2)} = t + V_e$, where V_e is a uniform random variable on the interval $(1/2, 3/2)$.

Remark 25 Figure 3.1(a) displays a random sample for these two models. The sample functional mean for model B1 is marked by circles and for model R1 by squares. Notice that models B1 and R1 differ in level when U_e takes value in $(0, 1/2)$ and V_e in $(1, 3/2)$ but they coincide when U_e and V_e take values in $(1/2, 1)$. This intersection causes a theoretical misclassification rate equal to 25% when the method *DFM0* is used. Moreover, the first derivative of models $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ coincides, so method *DFM1* will fail in this setting.

Simulation Exercise 2. We consider the following two functional data generating models:

Model B2. $\mathcal{X}_e^{(1)} = (t + U_e)^2$, where U_e is a uniform random variable on the interval $(0, 1)$.

Model R2. $\mathcal{X}_e^{(2)} = t^2 + V_e$, where V_e is a uniform random variable on the interval $(0, 1)$.

Remark 26 Figure 3.1(b) displays a random sample for these two models. The sample functional mean for model B2 is marked by circles and for model R2 by squares. Notice that models B2 and R2 generate functional observations that cross one another; but if we consider the first derivative, $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$, then they have significant level differences. The theoretical misclassification rate is equal to 12.5% when the method *DFM1* is used.

Simulation Exercise 3. We consider the following two functional data generating models:

Model B3. $\mathcal{X}_e^{(1)} = (t + U_e)^2 + 5/4$, where U_e is a uniform random variable on the interval $(0, 1)$.

Model R3. $\mathcal{X}_e^{(2)} = (t + V_e)^2$, where V_e is a uniform random variable on the interval $(1/2, 3/2)$.

Remark 27 Figure 3.1(c) displays a random sample for these two models. The sample functional mean for model B3 is marked by circles and for model R3 by squares. Notice that models B3 and R3 also generate functional observations that cross one another (the term $5/4$ in $\mathcal{X}^{(1)}$ is added in order to maximize the crossing) but if we consider the first derivatives, $D^1\chi^{(1)}(t)$ and $D^1\chi^{(2)}(t)$, then these have level differences in the same way as $\chi^{(1)}(t)$ and $\chi^{(2)}(t)$ generated by models B1 and R1, respectively. So, we have a theoretical misclassification rate equal to 25% when the method *DFM1* is used.

In figure 3.2 we present the results for the first simulation setting. Figure 3.2(a) gives the boxplots of the misclassification rate estimates for the four methods. As expected, the method *DFM0* has a misclassification rate of around 25% and the method *DFM1* is useless in this setting. Figures 3.2(b) and 3.2(c) give the boxplots of the estimated weights for methods *WI* and *WD*. Both methods give positive weights for the variable associated to $\chi^{(1)}$ and $\chi^{(2)}$ and zero weights for the variable associated to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$. Notice that in this case the variable $D^1\chi^{(1)} - D^1\chi^{(2)}$ has variance equal to zero since $D^1\chi_e^{(1)}(t) = D^1\chi_e^{(2)}(t) = 1$ for all e . In this simulation setting, methods *DFM0*, *WI* and *WD* have the same performance.

In figure 3.3, we present the results for the second simulation setting. Figure 3.3(a) gives the boxplots of the misclassification rate estimates for the four methods. In this case, method *DFM0* is outperformed by method *DFM1*, which obtains misclassification rates around the expected

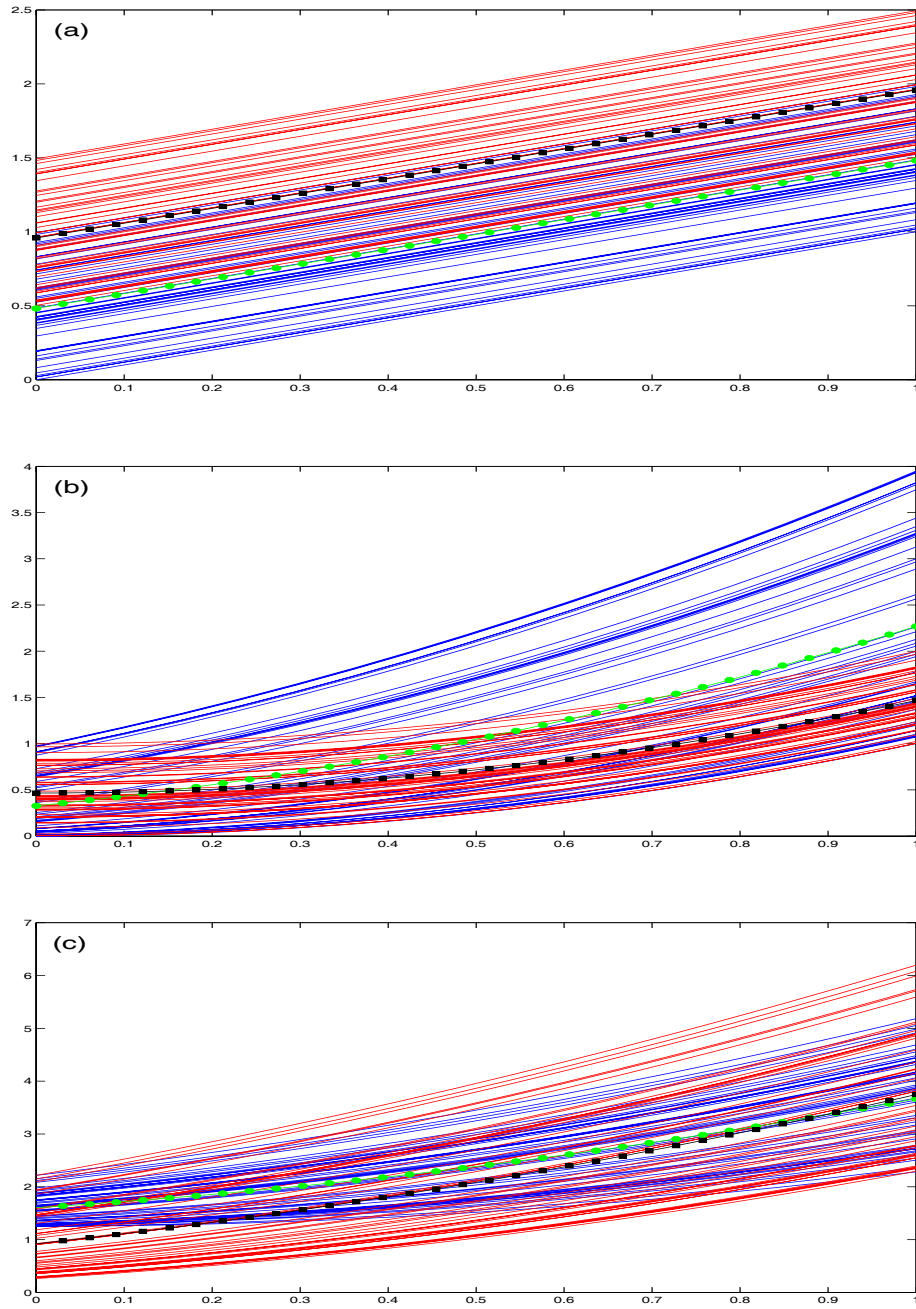


Figure 3.1: Plots of samples from the three simulation settings: (a) Functions following models B1 and R1; (b) Functions following models B2 and R2; (c) Functions following models B3 and R3.

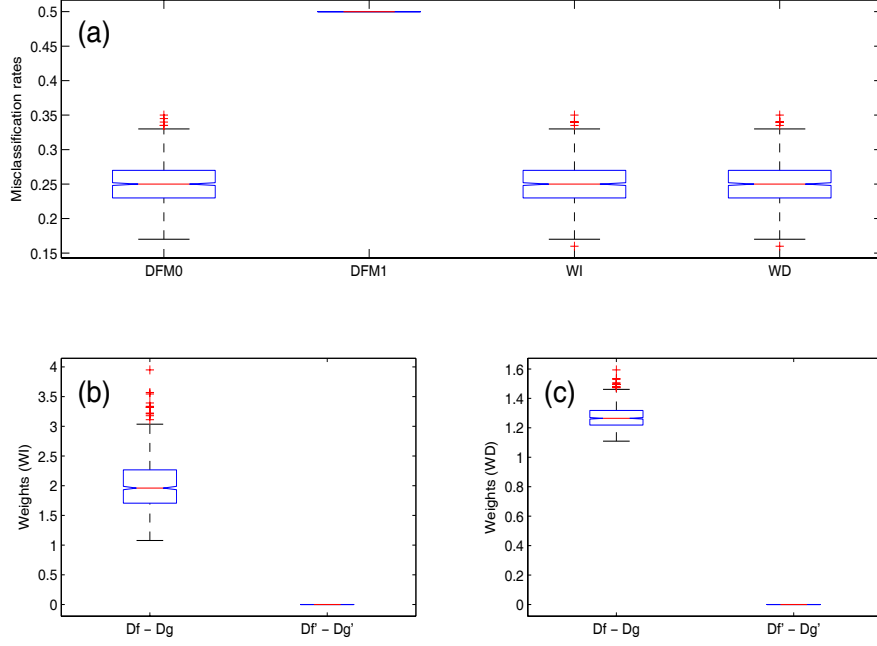


Figure 3.2: First simulation setting results: (a) Boxplots of the misclassification rates for methods $DFM0$, $DFM1$, WI and WD ; (b) Boxplots of the weights obtained for method WI ; (c) Boxplots of the weights obtained for method WD .

12.5%. Method WD has a performance similar to $DFM1$, and both are outperformed by method WI . Figures 3.3(b) and 3.3(c) give the boxplots of the estimated weights for methods WI and WD . In this case, method WI gives positive weights for the variable associated to $\chi^{(1)}$ and $\chi^{(2)}$ and negative (but higher in module) weights for the variable associated to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$, so the classification rule with WI is not based on a distance. Once we impose the positiveness on the weights, method WD gives positive weights for the variable associated to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$ and zero weights for the variable associated to $\chi^{(1)}$ and $\chi^{(2)}$. So, the classification rule with WD is a semidistance. In this setting and in the previous one, method WD selects the variable that has lower misclassification rates.

In figure 3.4, we present the results for the third simulation setting. Figure 3.4(a) gives the boxplots of the misclassification rate estimates for the four methods. In this case, method $DFM0$ is again outperformed by method $DFM1$, which obtains misclassification rates around the expected 25%. Both methods perform worse than the weighted procedures, WI and WD ; method WI has the best performance. Here, the improvement comes from the combination of variables associated to functions and their first derivatives. Figures 3.4(b) and 3.4(c) give the boxplots of the estimated

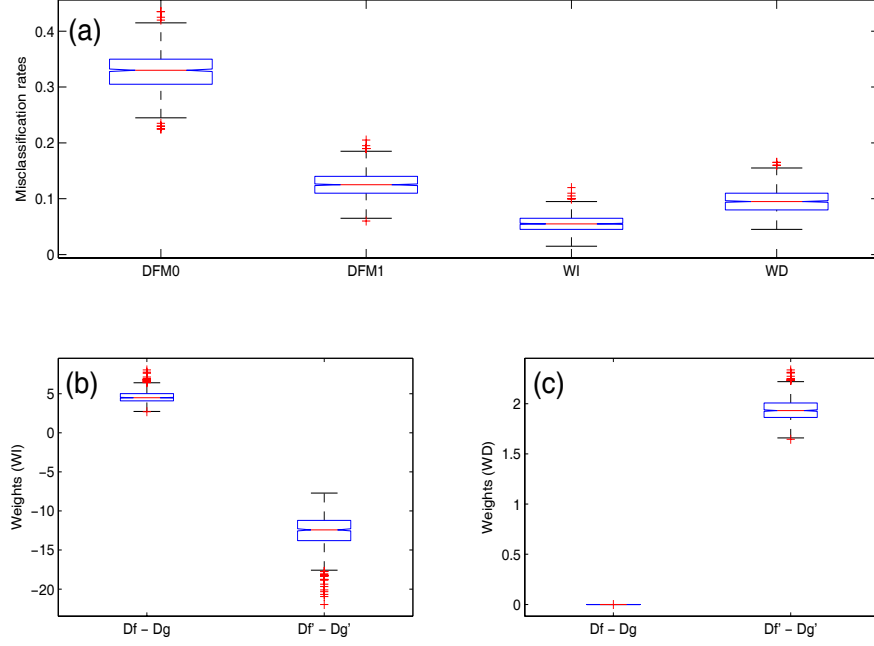


Figure 3.3: Second simulation setting results: (a) Boxplots of the misclassification rates for methods $DFM0$, $DFM1$, WI and WD ; (b) Boxplots of the weights obtained for method WI ; (c) Boxplots of the weights obtained for method WD .

weights for methods WI and WD . In this case, method WI gives positive weights for the variable associated to $\chi^{(1)}$ and $\chi^{(2)}$ in more than 25% of the replications and negative weights in the remaining ones. In all replications, WI gives negative weights for the variable associated to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$. For those replications where there are sign differences, the classification rule with WI is not a distance. This “inconvenience” is avoided by using the method WD . In this setting, the classification rule with WD is a semidistance in all cases and a distance in 75% of the replications.

3.4 Real Data Examples

In this section we illustrate the performance of our proposal in two benchmark data sets: (a) *Spectrometric data set*, consisting of 215 near-infrared spectra of meat samples obtained by a Tecator Infratec Food and Feed Analyzer; (b) *Growth curves data set*, consisting of the height (in centimeters) of 44 girls and 39 boys measured at a set of 31 ages from 1 to 18 years old.

In both examples, the original data was smoothed using a cubic smoothing spline with smoothing parameter equal to $1/(1 + h^3/6)$, where h is the average spacing of the data sites (see De Boor

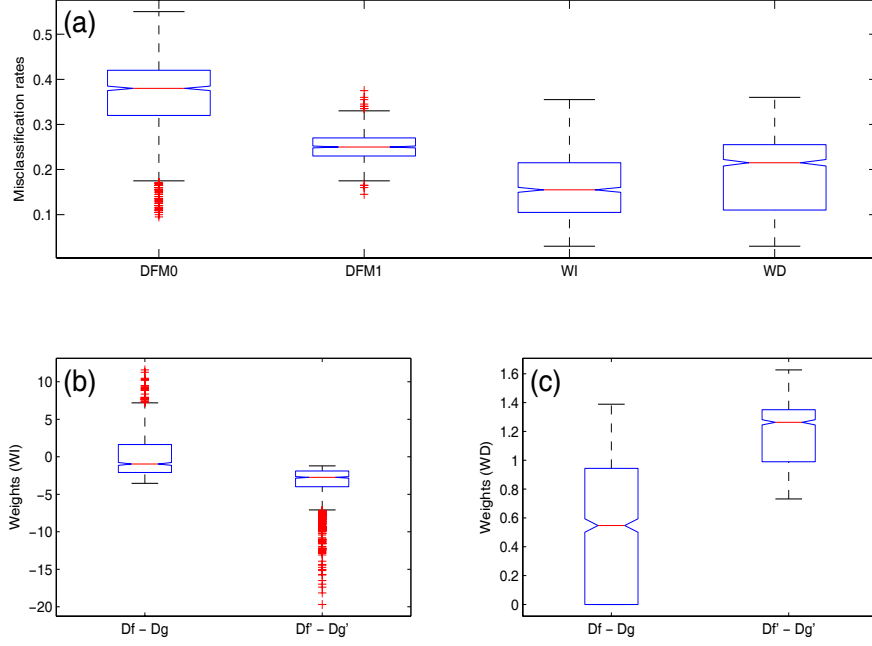


Figure 3.4: Third simulation setting results: (a) Boxplots of the misclassification rates for methods *DFM0*, *DFM1*, *WI* and *WD*; (b) Boxplots of the weights obtained for method *WI*; (c) Boxplots of the weights obtained for method *WD*.

[1978]).

In this section, the nomenclature for the different versions of the algorithm is that used in the previous section. Furthermore,

- *DFM2* denotes the classification with the distance to the sample functional mean calculated using the second derivatives of functions in the training set. That is, using the rule

$$\begin{cases} k = 1 & \text{if } x_3 < 0 \\ k = 2 & \text{otherwise} \end{cases}. \quad (3.36)$$

- Now the weighted approaches take into account up to the second derivative by considering

$$\mathbf{x} = (x_1, x_2, x_3)^t. \quad (3.37)$$

3.4.1 Spectrometric Data

The classification problem in the spectrometric data set consists in separating meat samples with a high fat content (more than 20%) from samples with low fat content (less than 20%). Among the

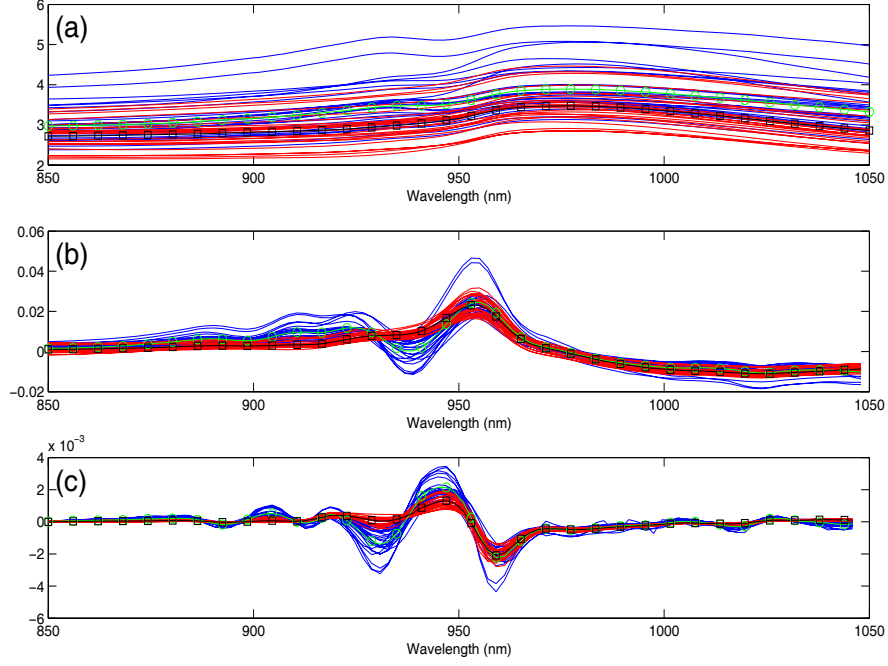


Figure 3.5: Sample from the spectrometric data set (wavelengths 850–1050 nm): (a) Data; (b) First derivative; (c) Second derivative.

215 samples, 77 have high fat content and 138 have low fat content. Figure 3.5 shows a sample of these 100-channel absorbance spectrum in the wavelength 850–1050 nm and the first and second derivatives.

Among others, Ferraty and Vieu (2003), Rossi and Villa (2006) and Li and Yu (2008) had considered the original spectrum and its derivatives for classification purpose and had concluded that the second derivative produces the lower misclassification rates.

In order to evaluate the performance of our proposal, we will split the data set into 120 spectra for training and 95 spectra for testing as in Rossi and Villa (2006) and Li and Yu (2008). The classification results shown in figure 3.6 are based on 1000 replications. Methods *WI* and *WD* obtain a mean misclassification rate equal to 2.02% and 2.32%, respectively. They improve the classification rules based on the second derivative, *DFM2*, which obtains 3.70%.

In this example, method *WI* gives positive weights to the variable associated with $\chi^{(1)}$ and $\chi^{(2)}$, and negative —but higher in module— weights for the variables associated with $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$ and with $D^2\chi^{(1)}$ and $D^2\chi^{(2)}$; so the classification rule with *WI* is not based on a distance. Method *WD* gives positive weights to the variables associated with $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$ and with $D^2\chi^{(1)}$ and $D^2\chi^{(2)}$, and zero weights for the variable associated with $\chi^{(1)}$ and $\chi^{(2)}$. Notice that

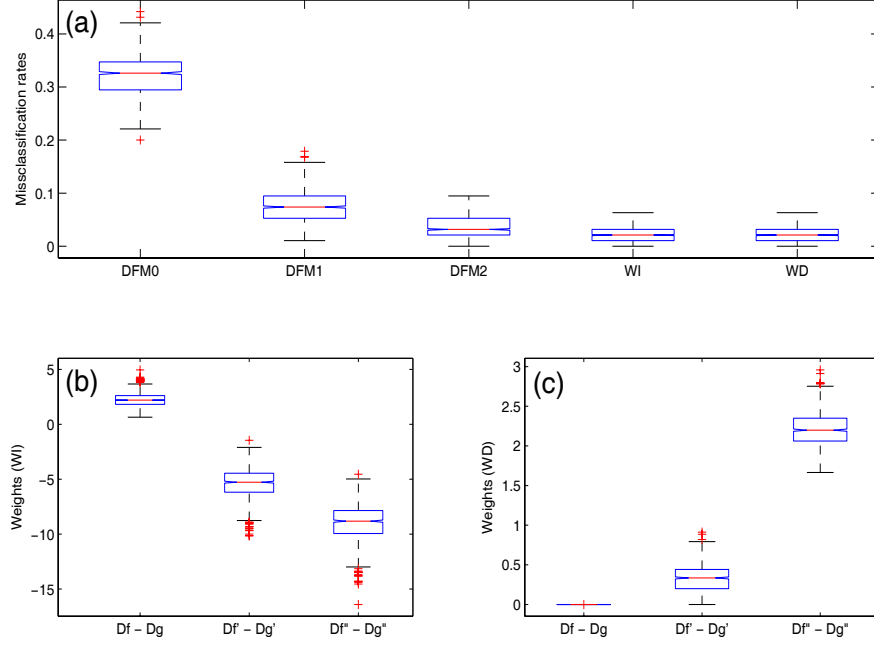


Figure 3.6: Spectrometric data set classification results: (a) Boxplots of the misclassification rates for methods $DFM0$, $DFM1$, WI and WD ; (b) Boxplots of the weights obtained for method WI ; (c) Boxplots of the weights obtained for method WD .

both procedures give the higher weights to the variable associated with $D^2\chi^{(1)}$ and $D^2\chi^{(2)}$, which is consistent with the results of Ferraty and Vieu (2003), Rossi and Villa (2006) and Li and Yu (2008).

The functional support vector machine proposed by Rossi and Villa (2006) obtains 3.28% (7.5%) using a linear (Gaussian) kernel and the spectra, and a 2.6% (3.28%) using a Gaussian (linear) kernel and the second derivative of the spectra.

The nonparametric functional method proposed by Ferraty and Vieu (2003) obtains a mean error of around 2% using the second derivative. Notice that Ferraty and Vieu (2003) use a training set with 160 spectra. In that setting, our mean misclassification rates are equal to 1.89% and 2.27%, respectively.

Li and Yu (2008) obtain 3.98%, 2.91% and 1.09% using the raw data, the first derivative and the second derivative, respectively. Notice that Li and Yu's method selects the data segments where the two populations have large differences, and then it combines the linear discriminant as a data reduction tool and the support vector machine as a classifier. These authors' method has three tuning parameters: number of segments, separation amongst segments and the SVM

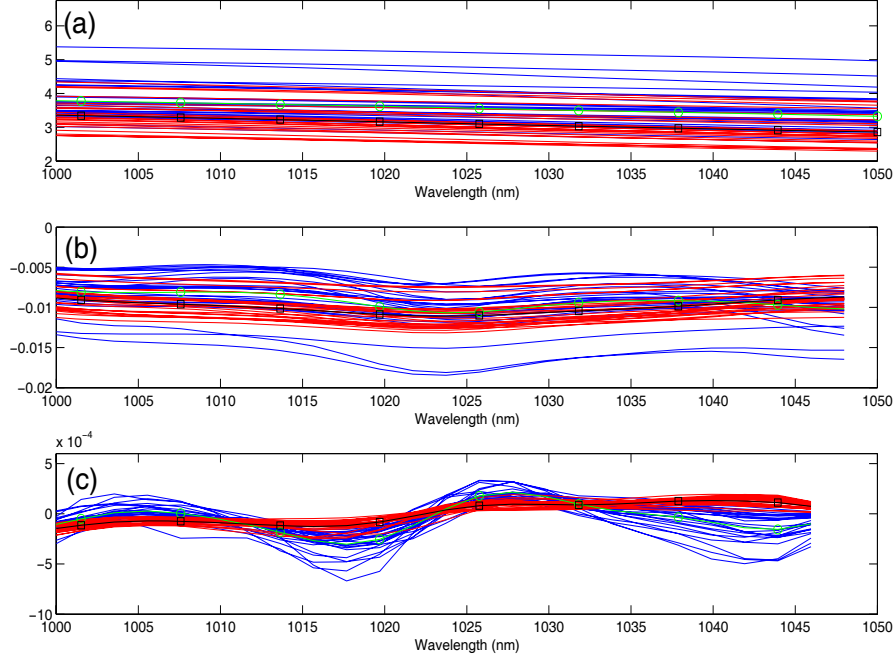


Figure 3.7: Sample from the spectrometric data set (wavelengths 1000–1050 nm): (a) Data; (b) First derivative; (c) Second derivative.

regularization parameter.

If we repeat our procedures using the channels in the wavelengths 1000–1050 nm, then we obtain 1.49% and 1.30%, using *WI* and *WD*, respectively. Figure 3.7 shows a sample of these spectrum in the wavelength 1000–1050 nm and the first and second derivatives. This segment, 1000–1050 nm, was obtained by cross-validation through a grid search. The design of a segmentation approach for selecting more than one segment is beyond the scope of this thesis and probably deserves separate research.

3.4.2 Growth Data

The classification problem in the growth data set consists of separating samples by sex, taking the growth curves as discriminant variables. Figure 3.8 shows a sample of these curves, measured in ages ranging from 1 to 18, and their first and second derivatives. López-Pintado and Romo (2006) had considered the growth curves (but not their derivatives) for classification purpose.

In order to evaluate the performance of our proposal, we will split the data set into 60 curves for training and the remaining 33 for testing. The classification results shown in figure 3.9 are based on 1000 replications. Weighted methods *WI* and *WD* have similar behaviour with means

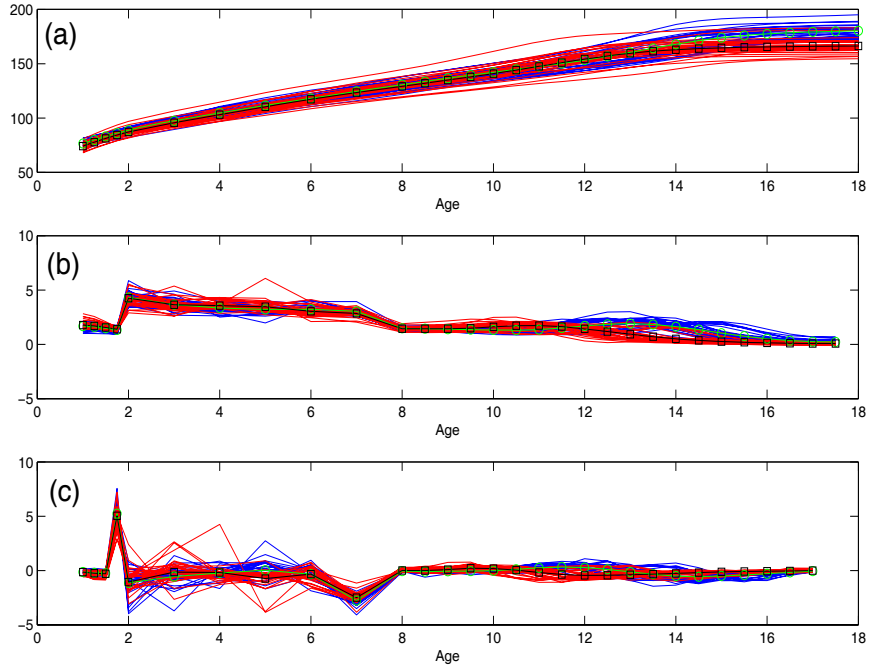


Figure 3.8: Sample from the growth data set: (a) Data; (b) First derivative; (c) Second derivative.

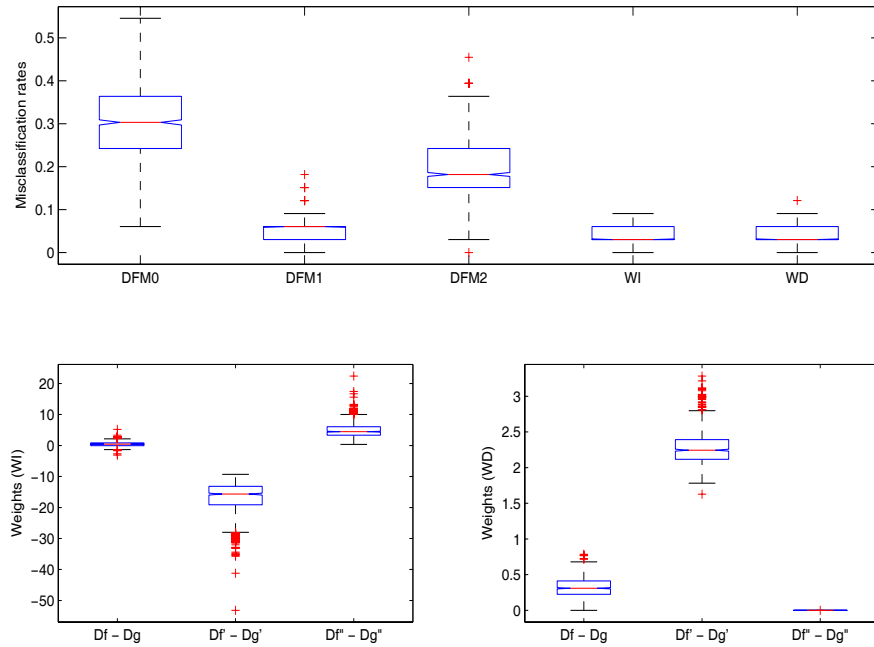


Figure 3.9: Growth data set classification results: (a) Boxplots of the misclassification rates for methods $DFM0$, $DFM1$, WI and WD ; (b) Boxplots of the weights obtained for method WI ; (c) Boxplots of the weights obtained for method WD .

misclassification rate equal to 3.65% and 3.75%, respectively. They improve the classification rules based on the raw data, on the first derivative or on the second derivative, which obtain 31.08%, 5.30% and 18.85%, respectively. The best result with the depth-based classification procedure proposed by López-Pintado and Romo (2006) was 14.86%.

In this example, method *WI* gives positive weights for the variable associated to $D^2\chi^{(1)}$ and $D^2\chi^{(2)}$ and negative (but higher in module) weights for the variables associated to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$; so the classification rule with *WI* is not based on a distance. Method *WD* gives positive weights for the variables associated to $\chi^{(1)}$ and $\chi^{(2)}$ and to $D^1\chi^{(1)}$ and $D^1\chi^{(2)}$; then, the classification rule is based on a distance.

3.5 Conclusions

In this chapter we have proposed a new approach for discriminating functional data. This method involves the use of distances to a representative function and its successive derivatives. Our simulation studies as well as our applications show that the method performs very well resulting in small training and testing classification errors. Applications to real data show that our procedure performs as well as —and in some cases better than— other classifications methods. In addition, our methodology provides, through the weights, information about the importance of each data transformation. Finally, some adaptability of our methodology to the different types of functional data can be achieved by selecting the distance $d(\cdot, \cdot)$ or the multivariate classification technique.

Chapter 4

Extensions and Further Work

Summary: In this chapter, possible generalisations and forthcoming work, for both classification methods, are itemised and outlined briefly.

There is much theory available on multivariate, time series and functional data, so techniques from the three frameworks can be applied, respectively, to the original time series and curves or to the constructed curves and variables: graphical representations, transformations, inference, typicality and robustness, etcetera. For example, graphical methods can be used for the detection of outliers, through the representation of the curves—in our first proposal—or our multivariate variables—in the second.

Nevertheless, a great complexity appears if we want to connect both theories, taking into account the fact that in our cases the functional or multivariate characters come, respectively, from original time series and functional data, with transformations and distance measurements.

4.1 Time Series Method

4.1.1 More than Two Populations

The generalisation to a K -group classification, with $K > 2$, is trivial. In the algorithms of sections 2.2.3 and 2.3.2, it is enough to consider the reference curves of the K groups and apply the classification rule

$$k = \operatorname{argmin}_{\{1, \dots, K\}} \{d(\chi(\lambda), \mathcal{R}^{(k)}(\lambda))\}. \quad (4.1)$$

with $\mathcal{R}^{(k)}(\lambda) = \bar{\chi}^{(k)}(\lambda)$ or $\mathcal{R}^{(k)}(\lambda) = \bar{\chi}^{\frac{\alpha}{\lambda}}{}^{(k)}(\lambda)$, respectively.

4.1.2 Clustering

From our first proposal it is concluded that the spectral distribution function contains information which is useful for the supervised classification of time series. The same information can be used for the unsupervised classification —or clustering— of time series, by tackling the corresponding/associated functional data problem in the frequency domain: some references on this subject are mentioned at the end of section 8.6 of Ramsay and Silverman (2006) and —with extension and in the nonparametric framework— in chapter 9 of Ferraty and Vieu (2006).

4.1.3 Other Depth Definitions

Other different definitions of depth can be considered, for example: Zuo and Serfling (2000), Vardi and Zhang (2000), Fraiman and Muniz (2001), Zuo (2003) and Cuevas et al. (2007).

4.2 Functional Data Method

4.2.1 Classical Assumptions

Some interesting questions are related to the fulfilment of the linear discriminant analysis assumptions and the behaviour when there are departures from them, possible corrections of these departures or extensions in order to deal with them. In general, the linear discriminant analysis, when used for classification, is quite robust to departures from the assumptions, especially for large sample sizes. In practice the multivariate techniques can be applied to our discriminant variables, but it would be very interesting to do some simulations studying the distribution of these variables for different functional models and distances.

The normality assumption can be tested, that is, how far from the normal distribution our variables x_1, \dots, x_p are. Under normality, the method becomes optimal; otherwise, functional or multivariate transformations to achieve normality could be considered. The normality of the discriminant variables implies the normality of the discriminant functions, so the rejection of this normality would imply the rejection of that normality. On the other hand, since the discriminant analysis is somewhat affected by the presence of atypical values, multivariate or functional

techniques could be used to leave these values out.

Equally important is testing the homoscedasticity assumption, that is, if $\Sigma_{\mathbf{x}}^{(k)} = \Sigma_{\mathbf{x}}$, $\forall k$, in order to combine or not the information in a unique pooled estimator of the covariance matrix. When the equality is rejected, quadratic —instead of linear— discriminant analysis should be used in our approach. The quadratic analysis, however, is more sensitive to the normality assumption. On the other hand, if the equality holds, the use of the quadratic version implies a loss of efficiency.

When it is not possible to reject the equality of the covariance matrices, it is important to test the equality of the means, that is, the hypothesis $\mu_{\mathbf{x}}^{(k)} = \mu_{\mathbf{x}} \quad \forall k$. If this equality cannot be rejected either, it is not possible to ensure from the samples that, in fact, there are two different underlying populations.

Finally, if high correlation were observed in the multivariate samples (induced by correlations in the functional samples), the linear discriminant analysis can be substituted by the *penalized discriminant analysis* of Hastie et al. (1995), a variant of the classical discriminant analysis specifically designed to cope with correlations.

4.2.2 Additional Constraint Embedding

The quadratic-form optimization problem arises frequently in literature. When there is a constraint, a known approach consists in finding an equivalent optimization problem without the constraint. This is done by embedding the constraint into the quadratic form; analitically, by introducing the constraint multipliers in the matrix of the quadratic form. For example, Jagannathan and Ma (2003) transform the minimum variance portfolio optimization problem

$$\mathbf{a} = \operatorname{argmin} \{ \mathbf{a}^t \mathbf{S} \mathbf{a} \} \quad \text{subject to} \quad \begin{cases} \sum_i a_i = 1 \\ 0 \leq a_i \leq \bar{a} \end{cases} \quad (4.2)$$

into the equivalent problem

$$\mathbf{a} = \operatorname{argmin} \{ \mathbf{a}^t \tilde{\mathbf{S}} \mathbf{a} \} \quad (4.3)$$

by considering $\tilde{\mathbf{S}} = \mathbf{S} + (\delta \mathbf{1}^t + \mathbf{1} \delta^t) - (\lambda \mathbf{1}^t + \mathbf{1} \lambda^t)$, where $\mathbf{1}$ is the column vector of ones.

For our optimization problem

$$\mathbf{a} = \operatorname{argmax} \left\{ \frac{\mathbf{a}^t \mathbf{B} \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} \right\} \quad \text{subject to} \quad \mathbf{a} \geq \mathbf{0} \quad [3.1]$$

we would be interested in defining new matrices $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{W}}$ so as to obtain the equivalent optimization problem

$$\mathbf{a} = \operatorname{argmax} \left\{ \frac{\mathbf{a}^t \tilde{\mathbf{B}} \mathbf{a}}{\mathbf{a}^t \tilde{\mathbf{W}} \mathbf{a}} \right\}. \quad (4.4)$$

This would allow using all the theory of the Fisher's discriminant analysis; for example, the expression of the discriminant function, when $\tilde{\mathbf{W}}$ is nonsingular, will be:

$$y = \mathbf{a}_p^t \mathbf{x} = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \tilde{\mathbf{W}}^{-1} \mathbf{x}. \quad (4.5)$$

More generally, for this unconstrained optimization problem general theory could be considered or developed, as that of McDonald (1979) and Kiers (1995) for quotients of quadratic forms. Thus, this constraint embedding problem has general theoretical interest.

4.2.3 Additional Constraint Avoidance

As we have mentioned in section 3.2.6, we introduced the nonnegativeness constraints for theoretical reasons. Perhaps the same theoretic advantage could sometimes be achieved through a different way. For example, by applying a transformation $\mathbf{t} : \mathbb{R}^p \rightarrow \mathbb{R}^p$ to the \mathbf{a}_F coefficients, instead of imposing the additional constraint. That is, by transforming the classical discriminant function in the following way:

$$\tilde{y} = \mathbf{t}(\mathbf{a}_F)^t \mathbf{x} = \sum_{i=1}^p t_i(\mathbf{a}_F) x_i, \quad (4.6)$$

where $\mathbf{t} = (t_1, \dots, t_p)^t$ with $t_i : \mathbb{R}^p \rightarrow \mathbb{R}$ such that $t_i(\mathbf{a}) \geq 0, \quad \forall i = 1, \dots, p$.

ADVANTAGES

It seems that this approach would add several advantages. On the one hand, the core optimization problem of the methodology would again be the classical Fisher's problem, so the usual interpretation of \mathbf{a}_F and the available software could be accessed. On the other hand, several transformations, $t(\cdot)$, could be tested without having to solve the optimization problem again, that is, without computing \mathbf{a}_F again.

DISADVANTAGES

Let us notice that the election of $\mathbf{t}(\cdot)$ is an ad hoc election, while the optimization problem—with the nonnegativeness restriction—is based on a general and transparent criterion.

SOME CASES

1. Isotropic: If $\mathbf{t} = (t, \dots, t)^t$
2. Anisotropic: If $\mathbf{t} = (t_1, \dots, t_p)^t$
3. Unidimensional components: If $\mathbf{t} = (t_1(a_1), \dots, t_p(a_p))^t$, that is, with $t_i : \mathbb{R} \rightarrow \mathbb{R}$.

THE CLASSIFICATION

In the previous cases, the classification rule could be

$$\begin{cases} k = 1 & \text{if } \tilde{y} > \mathbf{t}(\mathbf{a}_F)^t \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\ k = 2 & \text{otherwise} \end{cases}, \quad (4.7)$$

where the value $\mathbf{t}(\mathbf{a}_F)^t \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ would be the new cutoff point.

Remark 28 Under normality (of the multivariate variables), it is not possible that the classification with (4.7) would improve the classification with (B.38) of appendix B.4, as in this case $\mathbf{t}(\mathbf{a}_F)$ would have been probably found firstly instead of \mathbf{a}_F ; nevertheless, some improvement can be achieved when the normality does not hold. On the other hand, if the classification is not the aim of one type of methodology and the nonnegativeness is not desired for the same reasons than in our proposal, perhaps the classification can be done considering $\mathbf{t}(\mathbf{a}_F)$.

4.2.4 Transformation Importance

Our methodology provides information about the usefulness of the different function derivatives for classification purposes. In the same way, we could try to obtain information over data transformations different from derivation, that is, using the following discriminant variables instead of (3.26),

$$x_i = d(T_i(\chi), \overline{T_i(\chi^{(1)})}) - d(T_i(\chi), \overline{T_i(\chi^{(2)})}), \quad (4.8)$$

for $i = 1, 2, \dots, p$, where $\overline{T_i(\chi^{(k)})} = \frac{1}{n_k} \sum_{e=1}^{n_k} T_i(\chi_e^{(k)})$, $k = 1, 2$, with $T_i(\cdot)$ being an application between two functional spaces.

Notice that the information of different transformations could be combined (indirectly, through the distances) using an expression similar to (3.32), that is, with

$$y(\mathbf{x}) = \mathbf{a}^t \mathbf{x} = \sum_{i=1}^p a_i x_i$$

$$= \sum_{i=1}^p a_i d(T_i(\chi), \overline{T_i(\chi^{(1)})}) - \sum_{i=1}^p a_i d(T_i(\chi), \overline{T_i(\chi^{(2)})}). \quad (4.9)$$

4.2.5 Distance Importance

Given a data set, which are the important characteristics depends on the aim of the study for which data is being used. For example, a peak in smooth curves can indicate the presence of an event, and depending on whether or not there is interest in this event—or even in avoiding it—one particular distance should be used: $\|\chi\| = \max_{t \in I} \chi(t)$ or $\|\chi\| = \int_I \chi(t) dt$.

Then, similarly as in the previous subsection, the importance of several different distances can be studied using the variables

$$x_i = d_i(\chi, \overline{\chi^{(1)}}) - d_i(\chi, \overline{\chi^{(2)}}) \quad (4.10)$$

for $i = 1, 2, \dots, p$, where $d_i(\cdot, \cdot)$ are different distances and $\overline{\chi} = \frac{1}{n_k} \sum_{e=1}^{n_k} \chi_e^{(k)}$, $k = 1, 2$.

An important fact is that, again with an expression similar to (3.32), combinations of distances can be constructed:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{a}^t \mathbf{x} = \sum_{i=1}^p a_i x_i \\ &= \sum_{i=1}^p a_i d_i(\chi, \overline{\chi^{(1)}}) - \sum_{i=1}^p a_i d_i(\chi, \overline{\chi^{(2)}}). \end{aligned} \quad (4.11)$$

4.2.6 Several Discriminant Functions

The few discriminant variables considered by us, two or three, justified the use of only one discriminant function. Some of the previous extensions could require, however, several discriminant variables; for example, when many discriminant variables are considered or when more information needs to be captured. In this case, a multivariate compound variable \mathbf{y} has to be constructed in several consecutive steps, as in the classical discriminant analysis. Then, the classification rule becomes

$$k = \operatorname{argmin}_{\{1,2\}} \{d(\mathbf{y}(\mathbf{x}), \mathbf{y}(\overline{\mathbf{x}}^{(k)}))\}, \quad (4.12)$$

where now $d(\cdot, \cdot)$ is a distance in \mathbb{R}^q (here q is the length of \mathbf{y}).

4.2.7 Other Classification Methods

The use of the linear discriminant analysis in the algorithm (section 3.2.5) is not compulsory, and this technique can be substituted by another one. That is, once the multivariate variables are defined as in algorithm 3.2.5, they can be used for any other multivariate classification technique.

If the new method constructs a multivariate compound variable \mathbf{y} , the classification criterion can again be (4.12). Finally, if the weights in \mathbf{y} are nonnegative, our additional constraint would be unnecessary.

Conclusions

Although our proposals are not based on complex ideas, the theory included in the first chapter is necessary for a deep understanding of the second and third chapters. The different types of mathematical objects have been explained, including the basic statistical tools for inducing general information (models) from particular samples (data). At the end of this chapter, the problem of *classification* was presented in a way that the basic concepts of “property”, “distance” and “criterion” were highlighted.

In chapter two we have proposed a new frequency domain approach for time series classification based on the integrated periodograms of the series. When series are nonstationary, they are split into blocks and the integrated periodograms of the blocks are merged to construct a curve. This idea rests on the assumption that series are locally stationary; an example of definition of *locally stationary processes* has been given. Since the integrated periodogram is a function, the statistical tools for functional data analysis can be applied; concretely, the concept of *depth*, applied to functions, allows the use of a robust version of the functional mean. In our classification procedure new series are assigned to the class minimizing the distance between its corresponding curve and the group mean curve. Since the group mean can be affected by the presence of outliers, robustness of the classification method is achieved by substituting the mean curve with the α -trimmed mean, where for each group only the deepest elements are averaged. To evaluate our proposal in different scenarios, we have done simulation exercises containing several models and parameters, with both stationary and nonstationary series, as well as with different types of contamination. We have also illustrated the performance of our procedure in a real benchmark data set. Our proposal provides small error rates, robustness, and good computational behaviour, properties which make the methodology suitable for time series classification. The procedure suggests that the integrated periodogram contains useful information for classifying time series.

This previous work motivates both the functional data classification and the search of an

automatic procedure for selecting the derivative (or the crude functions) with the highest discriminant power. In chapter three we have proposed a new approach for discriminating functional data. This method involves the use of distances to a representative function and its successive derivatives. Since the information useful for discriminating is summarized in multivariate data, the appendix B is devoted to the most classical linear discriminant analysis. This method motivates the classificatory method and the introduction of an additional requirement that provides theoretical advantages. Our simulation studies, as well as our applications, show that the method performs very well, resulting in small training and testing classification errors. Applications to real data show that our procedure performs as well as —and in some cases better than— other classifications methods. Besides, our methodology provides, through the weights, information about the importance of each data transformation. Some adaptability of our methodology to the different types of functional data can be achieved by selecting the distance $d(\cdot, \cdot)$, the transformation $T(\cdot)$ or the multivariate classification technique providing y .

Finally, our proposals are easy to understand and quick. In Statistics there is not “the best” method, and ours have shown good behaviour in some schemes. In chapter four many possible extensions and further work has been outlined.

Appendix A

Vector Analysis

A Philological Note

Priestley (1981) makes the difference between the meanings of *multivariate* and *multidimensional* clear (see section 1.2 and the introduction of chapter 9):

Multivariate: “Of course, in a similar way we may have to consider the simultaneous variation of two, three, four, ..., or any number of related quantities, and such ‘collections’ of records are called *multivariate processes*”.

Multidimensional: “Generally, if a quantity depends on several variables it is termed a *multidimensional process*”.

In the same way, it would seem reasonable and coherent naming as:

$f : \mathbb{R} \rightarrow \mathbb{R}$	Univariate unidimensional function
$f : \mathbb{R}^p \rightarrow \mathbb{R}$	Univariate multidimensional function
$\mathbf{f} : \mathbb{R} \rightarrow \mathbb{R}^p$	Multivariate unidimensional function
$\mathbf{f} : \mathbb{R}^p \rightarrow \mathbb{R}^p$	Multivariate multidimensional function

Univariate Multidimensional Functions

Differentiation

Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be a differentiable function depending on the variables $\mathbf{x} = (x_1, \dots, x_p)^t$, then

Definition 33

$$\frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_p} \right)^t. \quad (\text{A.1})$$

Proposition 4 *If $f = \mathbf{x}^t \mathbf{c} = c_1 x_1 + \dots + c_p x_p$, where $\mathbf{c} = (c_1, \dots, c_p)^t$, then it holds that*

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^t \mathbf{c})}{\partial \mathbf{x}} = \mathbf{c} \quad (\text{A.2})$$

and, since $f^t = f$,

$$\frac{\partial f^t}{\partial \mathbf{x}} = \frac{\partial(\mathbf{c}^t \mathbf{x})}{\partial \mathbf{x}} = \mathbf{c}. \quad (\text{A.3})$$

Proof. See section 10 of Lütkepohl (1996). □

Proposition 5 *If $f = \mathbf{x}^t \mathbf{C} \mathbf{x}$, where \mathbf{C} is a square matrix, then it holds that*

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^t \mathbf{C} \mathbf{x})}{\partial \mathbf{x}} = (\mathbf{C} + \mathbf{C}^t) \mathbf{x} \quad (\text{A.4})$$

and, when \mathbf{C} is symmetric,

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^t \mathbf{C} \mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{C} \mathbf{x}. \quad (\text{A.5})$$

Proof. See section 10 of Lütkepohl (1996). □

Let be the differentiable functions $g : \mathbb{R}^p \rightarrow \mathbb{R}$ and $h : \mathbb{R}^p \rightarrow \mathbb{R}$. Then

Proposition 6 *If $f = g \cdot h$, then*

$$\frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial g}{\partial \mathbf{x}} h + g \frac{\partial h}{\partial \mathbf{x}} \right)^t \quad (\text{A.6})$$

and, as a particular case,

$$\frac{\partial f^2}{\partial \mathbf{x}} = \left(2f \frac{\partial f}{\partial \mathbf{x}} \right)^t. \quad (\text{A.7})$$

Proof. See section 10 of Lütkepohl (1996). □

Proposition 7 *If $f = g/h$, then*

$$\frac{\partial f}{\partial \mathbf{x}} = \left(\frac{1}{h^2} \left(\frac{\partial g}{\partial \mathbf{x}} h - g \frac{\partial h}{\partial \mathbf{x}} \right) \right)^t. \quad (\text{A.8})$$

Proof. See section 10 of Lütkepohl (1996). □

An equivalent definition, with different notation, is the following:

Definition 34 *The gradient of the function f is defined as the (column) vector*

$$\nabla f = \left(\frac{\partial f}{\partial \mathbf{x}} \right) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_p} \right)^t. \quad (\text{A.9})$$

When there are several multivariate variables with respect to which the gradient would be computed, a subindex will be added in the notation, for example, $\nabla_\beta f = \left(\frac{\partial f}{\partial \beta} \right)$.

Proposition 8 *When $\nabla f \neq \mathbf{0}$, it indicates —as a vector— the maximum variation direction of f .*

Proof. See section 2.5 of Marsden and Tromba (1991). □

Remark 29 This notation is preferred in Optimization theory (perhaps for the geometric interpretation of the gradient), while the previous, in terms of partial derivatives, is preferable in Multivariate analysis.

Definition 35 *The Hessian matrix of the function f is defined as*

$$\nabla^2 f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) \right)_{i,j}. \quad (\text{A.10})$$

Operators

The previous definitions can be rewritten in terms of the the following operators:

Definition 36 *The (vectorial) gradient operator (that applies on univariate multidimensional functions) is defined as*

$$\nabla = \left(\frac{\partial}{\partial \mathbf{x}} \right) = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_p} \right)^t. \quad (\text{A.11})$$

Definition 37 *The (scalar) Laplacian operator (that applies on univariate multidimensional functions) is defined as*

$$\nabla^2 = \nabla \cdot \nabla = \left(\frac{\partial}{\partial \mathbf{x}} \right)^t \left(\frac{\partial}{\partial \mathbf{x}} \right) = \sum_{i=1}^p \frac{\partial}{\partial x_i}, \quad (\text{A.12})$$

where the dot \cdot represents the canonical scalar product in \mathbb{R}^p .

Theoretical Results

Proposition 9 (Mean Value Theorem) *If $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is continuously differentiable over the linear segment from \mathbf{x} to \mathbf{y} , then an intermediate point in the segment exists, ξ , such that*

$$f(\mathbf{y}) - f(\mathbf{x}) = \nabla f(\xi)^t (\mathbf{y} - \mathbf{x}). \quad (\text{A.13})$$

Proposition 10 (Second Order Expansions) *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be twice continuously differentiable over an open sphere \mathcal{S} centered at a vector \mathbf{x} . Then*

(a) *For all \mathbf{y} such that $\mathbf{x} + \mathbf{y} \in \mathcal{S}$,*

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^t \nabla f(\mathbf{x}) + \frac{1}{2} \mathbf{y}^t \left(\int_0^1 \left(\int_0^t \nabla^2 f(\mathbf{x} + \tau \mathbf{y}) d\tau \right) dt \right) \mathbf{y}. \quad (\text{A.14})$$

(b) *For all \mathbf{y} such that $\mathbf{x} + \mathbf{y} \in \mathcal{S}$, there exists an $\alpha \in [0, 1]$ such that*

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^t \nabla f(\mathbf{x}) + \frac{1}{2} \mathbf{y}^t \nabla^2 f(\mathbf{x} + \alpha \mathbf{y}) \mathbf{y}. \quad (\text{A.15})$$

(c) *For all \mathbf{y} such that $\mathbf{x} + \mathbf{y} \in \mathcal{S}$, there holds*

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + \mathbf{y}^t \nabla f(\mathbf{x}) + \frac{1}{2} \mathbf{y}^t \nabla^2 f(\mathbf{x}) \mathbf{y} + o(\|\mathbf{y}\|^2). \quad (\text{A.16})$$

Proof.

See appendix A.5 of Bertsekas (1999).

□

Multivariate Multidimensional Functions

Let $\mathbf{f} : \mathbb{R}^p \rightarrow \mathbb{R}^p$ be a differentiable function, say $\mathbf{f} = (f_1, \dots, f_{m_f})^t$.

Differentiation

Definition 38 *The gradient matrix of a multivariate multidimensional function \mathbf{f} is defined as the matrix whose columns are ∇f_j , that is,*

$$\nabla \mathbf{f} = (\nabla f_1 \cdots \nabla f_{m_f})_j = \left(\left(\frac{\partial f_1}{\partial \mathbf{x}} \right) \cdots \left(\frac{\partial f_{m_f}}{\partial \mathbf{x}} \right) \right)_j. \quad (\text{A.17})$$

Remark 30 Instead of this definition, some *matricial differentiation* can be defined as the *vectorial differentiation* of section A.

Definition 39 *The Jacobian of a function is defined as the transpose of the gradient.*

Operators

Definition 40 *The (scalar) divergence operator (that applies on multivariate multidimensional functions) could be defined as*

$$\nabla \cdot = \left(\frac{\partial}{\partial \mathbf{x}} \right) \cdot = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_p} \right)^t \cdot, \quad (\text{A.18})$$

where \cdot represents the canonical scalar product in \mathbb{R}^p .

Definition 41 *The (vectorial) rotational operator (that applies on multivariate multidimensional functions) could be defined as*

$$\nabla \times = \left(\frac{\partial}{\partial \mathbf{x}} \right) \times = \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_p} \right)^t \times, \quad (\text{A.19})$$

where \times represents the canonical vectorial product in \mathbb{R}^p .

With these operators, the following could be defined:

Definition 42 *The (vectorial) divergence of \mathbf{f} is defined as*

$$\text{div}(\mathbf{f}) = \nabla^t \cdot \mathbf{f}. \quad (\text{A.20})$$

Definition 43 *The (vectorial) rotational of \mathbf{f} is defined as*

$$\text{rot}(\mathbf{f}) = \nabla \times \mathbf{f}. \quad (\text{A.21})$$

Remark 31 Definitions, applications and the interpretation can be found in sections 3.3, 3.4 and 8.4 of Marsden and Tromba (1991).

Appendix B

Fisher's Discriminant Analysis

In this appendix, the linear discriminant analysis is presented in the original Fisher's (1936) form, in the sense of not assuming either normality or equality of the group covariances; that is, nonparametrically and with possible heteroscedasticity. These two assumptions were introduced later also under the denomination *linear discriminant analysis*. This classification method is defined as an optimization problem from matrices that expresses the sample variability information. The analytical solution, the geometrical interpretation and the assignment of new elements to a population are explained.

B.1 Motivation

Let $\mathbf{X} = (X_1, \dots, X_p)^t$ be a random vector with mean $\mu_{\mathbf{X}} = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_p))^t$ and covariance matrix $\Sigma_{\mathbf{X}} = (\sigma_{ij})_{i,j} = (\text{cov}(X_i, X_j))_{i,j} = \mathbb{E}((\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{X} - \mu_{\mathbf{X}})^t)$; when there are $P^{(k)}$, $k = 1, \dots, K$, populations where the vector distribution is $\mathbf{X}^{(k)} = (X_1^{(k)}, \dots, X_p^{(k)})^t$, with parameters $\mu_{\mathbf{X}}^{(k)}$ and $\Sigma_{\mathbf{X}}^{(k)}$, respectively, capturing the differences between the groups from the distribution of \mathbf{X} is a subject of great interest.

On the other hand, it is frequently convenient or necessary to summarize the information of a vector in a shorter one; that is, to consider $\mathbf{Y} = (Y_1, \dots, Y_q)^t$, with $q < p$, instead of $\mathbf{X} = (X_1, \dots, X_p)^t$.

The previous two tasks can be done simultaneously via the following multiple transformation, where the coefficients can be interpreted as weights (in the sense explained in section B.2.3):

$$Y_j^{(k)} = a_{j1}X_1^{(k)} + \dots + a_{jp}X_p^{(k)}, \quad j = 1, \dots, q \quad (\text{B.1})$$

or, in matrix notation,

$$\mathbf{Y}^{(k)} = \mathbf{A}^t \mathbf{X}^{(k)}, \quad (\text{B.2})$$

where $\mathbf{A} = (a_{ij})$ is the $p \times q$ matrix of the coefficients. Notice that \mathbf{A} is independent of k , that is, independent of the population. The superscript $^{(k)}$ has been maintained in the notation to highlight that the new vector \mathbf{Y} also has a different distribution in each population, and that the election of \mathbf{A} must preserve or increase this difference so that \mathbf{Y} is suitable for discrimination. The covariance matrix of $\mathbf{Y}^{(k)}$ is $\Sigma_{\mathbf{Y}}^{(k)} = \mathbf{A}^t \Sigma_{\mathbf{X}}^{(k)} \mathbf{A}$.

When the model-vs-datum approach is applied, usually the unknown theoretical information must be inferred from samples. Let us consider, for each population k , the sample

$$(\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_{n_k}^{(k)}) = \begin{pmatrix} x_{11}^{(k)} & \cdots & x_{1n_k}^{(k)} \\ \vdots & \ddots & \vdots \\ x_{p1}^{(k)} & \cdots & x_{pn_k}^{(k)} \end{pmatrix}, \quad k = 1, \dots, K \quad (\text{B.3})$$

where $\mathbf{x}_j^{(k)}$, the j -th column of the matrix, contains the j -th element of the sample, and $n = \sum_{k=1}^K n_k$. In the following subsections some known theory of this multivariate framework is given in order to motivate Fisher's method criterion.

B.1.1 Parameter Estimation

The parameters of the distributions can be estimated—for each population—as explained in section 1.2.3. In addition, when homoscedasticity hypothesis holds, $\Sigma_{\mathbf{X}}^{(k)} = \Sigma_{\mathbf{X}}$, $k = 1, \dots, K$, the sample information can be combined and the matrix $\Sigma_{\mathbf{X}}$ is estimated with bias by $\hat{\Sigma}_{\mathbf{X}} = \sum_{k=1}^K \frac{n_k}{n} \hat{\Sigma}_{\mathbf{X}}^{(k)}$ and without bias by $\mathbf{S}_{\mathbf{X}} = \frac{n}{n-K} \hat{\Sigma}_{\mathbf{X}}$. See remark 33 below for an interpretation of these estimations.

B.1.2 Variability Information

Information about the within- and between-group variabilities are provided, respectively, by the *within-class scatter matrix*

$$\mathbf{W} = \sum_{k=1}^K \sum_{e=1}^{n_k} (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)}) (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)})^t \quad (\text{B.4})$$

and the *between-class scatter matrix*

$$\mathbf{B} = \sum_{k=1}^K n_k (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^t, \quad (\text{B.5})$$

where $\bar{\mathbf{x}} = n^{-1} \sum_{k=1}^K n_k \bar{\mathbf{x}}^{(k)}$ is the global mean. The *total scatter matrix*,

$$\mathbf{T} = \sum_{k=1}^K \sum_{e=1}^{n_k} (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}})(\mathbf{x}_e^{(k)} - \bar{\mathbf{x}})^t, \quad (\text{B.6})$$

expresses the total variability and is the sum of the previous quantities, $\mathbf{W} + \mathbf{B} = \mathbf{T}$.

These three matrices are, by definition, symmetric and positive semidefinite. In addition, \mathbf{W} is usually nonsingular (full rank) if $n > p$, and, as a consequence, positive definite.

Remark 32 The discriminant analysis is a supervised classification technique where the membership information is exploited through these variability matrices.

Remark 33 An important observation is that from definition (B.4) it holds that

$$\mathbf{W} = \sum_{k=1}^K n_k \hat{\Sigma}_{\mathbf{x}}^{(k)} = (n - K) \mathbf{S}_{\mathbf{x}}. \quad (\text{B.7})$$

This implies that both matrices, \mathbf{W} and $\mathbf{S}_{\mathbf{x}}$, could be used in the statements of this appendix. A positive constant factor does not change the optimization problems that will be considered. Nevertheless, we shall use \mathbf{W} since it maintains its meaning as variability matrix, while the matrix $\mathbf{S}_{\mathbf{x}}$ makes sense as an estimator only under the fulfilment of the equal group variability assumption (homoscedasticity).

COMPOUND VARIABLES

Given a compound variable $\mathbf{y} = \mathbf{A}^t \mathbf{x}$, it is important to establish the relation between the scatter matrices of \mathbf{y} —say $\mathbf{W}_{\mathbf{y}}$, $\mathbf{B}_{\mathbf{y}}$ and $\mathbf{T}_{\mathbf{y}}$ — and those of \mathbf{x} —say $\mathbf{W}_{\mathbf{x}}$, $\mathbf{B}_{\mathbf{x}}$ and $\mathbf{T}_{\mathbf{x}}$.

Proposition 11 *If $\mathbf{y} = \mathbf{A}^t \mathbf{x}$, then*

$$1. \mathbf{W}_{\mathbf{y}} = \mathbf{A}^t \mathbf{W}_{\mathbf{x}} \mathbf{A}.$$

$$2. \mathbf{B}_{\mathbf{y}} = \mathbf{A}^t \mathbf{B}_{\mathbf{x}} \mathbf{A}.$$

$$3. \mathbf{T}_{\mathbf{y}} = \mathbf{A}^t \mathbf{T}_{\mathbf{x}} \mathbf{A}.$$

Proof. Let $\mathbf{y}_e^{(k)}$ be the e -th element of the k -th sample; then

$$\mathbf{W}_{\mathbf{y}} = \sum_{k=1}^K \sum_{e=1}^{n_k} (\mathbf{y}_e^{(k)} - \bar{\mathbf{y}}^{(k)})(\mathbf{y}_e^{(k)} - \bar{\mathbf{y}}^{(k)})^t$$

$$\begin{aligned}
&= \sum_{k=1}^K \sum_{e=1}^{n_k} (\mathbf{A}^t \mathbf{x}_e^{(k)} - \mathbf{A}^t \bar{\mathbf{x}}^{(k)}) (\mathbf{A}^t \mathbf{x}_e^{(k)} - \mathbf{A}^t \bar{\mathbf{x}}^{(k)})^t \\
&= \sum_{k=1}^K \sum_{e=1}^{n_k} \mathbf{A}^t (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)}) (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)})^t \mathbf{A} \\
&= \mathbf{A}^t \left(\sum_{k=1}^K \sum_{e=1}^{n_k} (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)}) (\mathbf{x}_e^{(k)} - \bar{\mathbf{x}}^{(k)})^t \right) \mathbf{A} = \mathbf{A}^t \mathbf{W}_x \mathbf{A}, \tag{B.8}
\end{aligned}$$

$$\begin{aligned}
\mathbf{B}_y &= \sum_{k=1}^K n_k (\bar{\mathbf{y}}^{(k)} - \bar{\mathbf{y}}) (\bar{\mathbf{y}}^{(k)} - \bar{\mathbf{y}})^t \\
&= \sum_{k=1}^K n_k (\mathbf{A}^t \bar{\mathbf{x}}^{(k)} - \mathbf{A}^t \bar{\mathbf{x}}) (\mathbf{A}^t \bar{\mathbf{x}}^{(k)} - \mathbf{A}^t \bar{\mathbf{x}})^t \\
&= \sum_{k=1}^K n_k \mathbf{A}^t (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^t \mathbf{A} \\
&= \mathbf{A}^t \left(\sum_{k=1}^K n_k (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}}) (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^t \right) \mathbf{A} = \mathbf{A}^t \mathbf{B}_x \mathbf{A}, \tag{B.9}
\end{aligned}$$

and

$$\mathbf{T}_y = \mathbf{W}_y + \mathbf{B}_y = \mathbf{A}^t \mathbf{W}_x \mathbf{A} + \mathbf{A}^t \mathbf{B}_x \mathbf{A} = \mathbf{A}^t (\mathbf{W}_x + \mathbf{B}_x) \mathbf{A} = \mathbf{A}^t \mathbf{T}_x \mathbf{A}. \tag{B.10}$$

□

POSITIVENESS

The previous computes include a proof of the following statement.

Proposition 12 *Given a multivariate variable \mathbf{x} , its scatter matrices \mathbf{W} , \mathbf{B} and \mathbf{T} are positive semidefinite.*

Proof. It can be written

$$\mathbf{y}_e^{(k)} = (y_{e,1}^{(k)}, \dots, y_{e,q}^{(k)}) \tag{B.11}$$

$$\bar{\mathbf{y}}^{(k)} = \frac{1}{n_k} \sum_{e=1}^{n_k} \mathbf{y}_e^{(k)} = \left(\frac{1}{n_k} \sum_{e=1}^{n_k} y_{e,1}^{(k)}, \dots, \frac{1}{n_k} \sum_{e=1}^{n_k} y_{e,q}^{(k)} \right) \tag{B.12}$$

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{k=1}^K n_k \bar{\mathbf{y}}^{(k)} = \left(\frac{1}{n} \sum_{k=1}^K \sum_{e=1}^{n_k} y_{e,1}^{(k)}, \dots, \frac{1}{n} \sum_{k=1}^K \sum_{e=1}^{n_k} y_{e,q}^{(k)} \right). \tag{B.13}$$

The positiveness is obtained by looking to both (B.8) and (B.9) from the bottom up and considering the univariate case $y = \mathbf{a}^t \mathbf{x}$, that is:

$$\begin{aligned} \mathbf{a}^t \mathbf{W} \mathbf{a} &= \sum_{k=1}^K \sum_{e=1}^{n_k} (y_e^{(k)} - \bar{y}^{(k)})(y_e^{(k)} - \bar{y}^{(k)})^t \\ &= \sum_{k=1}^K \sum_{e=1}^{n_k} \left(y_e^{(k)} - \frac{1}{n_k} \sum_{e=1}^{n_k} y_e^{(k)} \right)^2 \geq 0 \end{aligned} \quad (\text{B.14})$$

and

$$\begin{aligned} \mathbf{a}^t \mathbf{B} \mathbf{a} &= \sum_{k=1}^K n_k (\bar{y}^{(k)} - \bar{y})(\bar{y}^{(k)} - \bar{y})^t \\ &= \sum_{k=1}^K n_k \left(\frac{1}{n_k} \sum_{e=1}^{n_k} y_e^{(k)} - \frac{1}{n} \sum_{k=1}^K \sum_{e=1}^{n_k} y_e^{(k)} \right)^2 \geq 0. \end{aligned} \quad (\text{B.15})$$

Finally, the matrix \mathbf{T} is positive semidefinite as a consequence of being the sum of positive semidefinite matrices.

□

B.1.3 Splitting Criterion

Given a unique sample with elements of both populations, minimizing a functional of \mathbf{W} or maximizing a functional of \mathbf{B} is a reasonable criterion for splitting the sample from the information provided by the vector \mathbf{x} . Many techniques are based on this idea.

B.1.4 Case $q = 1$: One Function

Some methods in literature choose —with different criteria— the linear combinations $y_j^{(k)}$ one at a time; thus, the case $q = 1$ is specially considered:

$$y^{(k)} = a_1 x_1^{(k)} + \dots + a_p x_p^{(k)} = \mathbf{a}^t \mathbf{x}^{(k)}, \quad k = 1, 2, \dots, K, \quad (\text{B.16})$$

with $\mathbf{a} = (a_1, \dots, a_p)^t$. For this new compound variable, $\bar{y}^{(k)} = \mathbf{a}^t \bar{\mathbf{x}}^{(k)}$ and $(s_y^{(k)})^2 = \mathbf{a}^t \mathbf{S}_{\mathbf{x}}^{(k)} \mathbf{a}$, where $\mathbf{S}_{\mathbf{x}}^{(k)}$ is the within-group sample covariance matrix. For classifying purposes, the variable $y^{(k)}$ must discriminate as much as possible. Following the idea of the above-mentioned splitting criterion, the interest is in finding \mathbf{a} so as to minimize the within-group dispersion, $W_y = \mathbf{a}^t \mathbf{W} \mathbf{a}$, or to maximize the between-group dispersion, $B_y = \mathbf{a}^t \mathbf{B} \mathbf{a}$. Then, the choice of \mathbf{a} can be formulated as an *optimal weighting problem*.

B.2 The Optimization Problem

A general fraction increases either when the numerator increases or the denominator decreases; that is, two criteria are combined and taken into account at the same time. Fisher's proposal is based on a trade-off criterion between maximizing B_y and minimizing W_y , as it maximizes the —sometimes termed— *generalised Rayleigh quotient*: $\lambda = B_y/W_y$ (this greek letter has no relation with the λ used in chapter 2). This quantity makes —mathematically— sense only when $W_y > 0$, that is, when \mathbf{W} is positive definite, or, equivalently in this case, is a nonsingular matrix: then, this hypothesis is necessary throughout this appendix and chapter 3. To add each consecutive compound function, this method also maximizes this quantity but with the imposition of uncorrelation with the previous combination.

Of special interest is the case of one *discriminant function* y and several *discriminant variables* x_1, \dots, x_p , consists in finding $\mathbf{a} \in \mathbb{R}^p$ such that:

$$\mathbf{a} = \operatorname{argmax} \{\lambda(\mathbf{a})\} = \operatorname{argmax} \left\{ \frac{B_y}{W_y} \right\} = \operatorname{argmax} \left\{ \frac{\mathbf{a}^t \mathbf{B} \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} \right\}. \quad (\text{B.17})$$

This is an unconstrained ($\mathcal{D} = \mathbb{R}^p$) *nonlinear (quadratic) optimization problem*.

The analytical resolution of the problem is obtained by solving

$$\mathbf{0} = \frac{\partial \lambda}{\partial \mathbf{a}} = \frac{2[\mathbf{B} \mathbf{a}(\mathbf{a}^t \mathbf{W} \mathbf{a}) - (\mathbf{a}^t \mathbf{B} \mathbf{a}) \mathbf{W} \mathbf{a}]}{(\mathbf{a}^t \mathbf{W} \mathbf{a})^2} = \frac{2[\mathbf{B} \mathbf{a} - \lambda \mathbf{W} \mathbf{a}]}{\mathbf{a}^t \mathbf{W} \mathbf{a}}, \quad (\text{B.18})$$

so the expression

$$\mathbf{0} = \mathbf{B} \mathbf{a} - \lambda \mathbf{W} \mathbf{a} = (\mathbf{B} - \lambda \mathbf{W}) \mathbf{a} \quad (\text{B.19})$$

is the eigenequation of the problem. For a nonnull solution to exist, it is necessary that $|\mathbf{B} - \lambda \mathbf{W}| = 0$. As \mathbf{W} is invertible (nonsingular) by hypothesis,

$$\mathbf{0} = \mathbf{W}^{-1}(\mathbf{B} - \lambda \mathbf{W}) \mathbf{a} = (\mathbf{W}^{-1} \mathbf{B} - \lambda \mathbf{I}) \mathbf{a}. \quad (\text{B.20})$$

The interest is in the largest eigenvalue of the matrix $\mathbf{W}^{-1} \mathbf{B}$.

SET OF SOLUTIONS

As $\lambda(c\mathbf{a}) = \lambda(\mathbf{a}) \quad \forall c \in \mathbb{R}, c \neq 0$, that is, $\lambda(\mathbf{a})$ is *homogeneous*, the solution —it exists— will not be a unique vector but an infinite family of them, denoted by

$$V_{\mathbf{a}}^* = \{c\mathbf{a} \mid c \in \mathbb{R}, c \neq 0\}. \quad (\text{B.21})$$

If \mathbf{a} is a nonnull eigenvector of λ , so is any element of $V_{\mathbf{a}}^*$; that is, the set of eigenvectors is the solution to (B.20): mathematically it is necessary to include the null vector as solution of the equation, but this vector makes no sense in the optimization problem. Let us denote the solution of this optimization problem by the pair $(V_{\mathbf{a}_F}^*, \lambda_F)$.

ANOTHER INTERPRETATION

Another interesting interpretation arises when the generalised Rayleigh quotient is written as

$$\lambda(\mathbf{a}) = \frac{B_y}{W_y} = \frac{\mathbf{a}^t \mathbf{B} \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} = \frac{\mathbf{a}^t (\mathbf{B} + \mathbf{W} - \mathbf{W}) \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} = \frac{\mathbf{a}^t \mathbf{T} \mathbf{a}}{\mathbf{a}^t \mathbf{W} \mathbf{a}} - 1. \quad (\text{B.22})$$

This decomposition shows that the previous maximization problem can be interpreted as maximizing the total variability while minimizing the within-class variability.

UNDER NORMALITY

As a final and general theoretical comment, when the distributions of $\mathbf{X}^{(k)}$ are normal, Fisher's approach is optimal in the sense of minimizing the misclassification probability.

B.2.1 Equivalent Problems

Usually the previous optimization problem is solved via equivalent constrained problems.

INEQUALITY CONSTRAINED PROBLEM

For fixed \mathbf{a} , since $\|c\mathbf{a}\| = |c|\|\mathbf{a}\|$ the scalar c can always be chosen so that the set of solutions $V_{\mathbf{a}}^*$ has a “representant” as close to the origin $\mathbf{0}$ as desired. This means that the search can be restricted to any domain \mathcal{D} having the origin as an interior point; concretely, the compact $(p-1)$ -dimensional sphere $\{\mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\}$ can be considered.

On the other hand, as \mathbf{W} is positive semidefinite, by proposition 13.d (in appendix C), the function $h(\mathbf{a}) = \mathbf{a}^t \mathbf{W} \mathbf{a}$ is convex; thus, from proposition 14.c (in appendix C) the set $\{\mathbf{a} \in \mathbb{R}^p \mid h(\mathbf{a}) = \mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\}$ is convex; finally, these facts and proposition 15 provide the equivalence—under existence—of the following optimization problem:

$$\mathbf{a} = \operatorname{argmax} \{\mathbf{a}^t \mathbf{B} \mathbf{a}\} \quad \text{subject to} \quad \mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1, \quad (\text{B.23})$$

where the feasible domain is $\mathcal{D} = \{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\}$.

For fixed λ , the solutions \mathbf{a} can be interpreted geometrically in \mathbb{R}^p as the intersection of a one dimensional linear subspace and the volume (sphere) whose frontier is $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$, that is $V_{\mathbf{a}} \cap \{\mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\}$ (strictly, the point $\mathbf{a} = \mathbf{0}$ must be excluded).

EQUALITY CONSTRAINED PROBLEM

The original optimization problem and the previous one, due to B.21, do not guarantee the uniqueness of the solution. To avoid the arbitrary scale factor and obtain a unique solution, usually the constraint is added in the form:

$$\mathbf{a} = \operatorname{argmax} \{\mathbf{a}^t \mathbf{B} \mathbf{a}\} \quad \text{subject to} \quad \mathbf{a}^t \mathbf{W} \mathbf{a} = 1, \quad (\text{B.24})$$

whose solution is $(\mathbf{a}_F, \lambda_F)$, with $\mathbf{a}_F^t \mathbf{W} \mathbf{a}_F = 1$ and $\lambda_F = \mathbf{a}_F^t \mathbf{B} \mathbf{a}_F$. Now, the feasible region is $\mathcal{D} = \{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a}^t \mathbf{W} \mathbf{a} = 1\}$. The calculations leading to the explicit expression of \mathbf{a}_F are given in section B.3. On the one hand, the constraint is only fixing the value of the constant c (defined in B.21), and, on the other hand, $V_{\mathbf{a}}^*$ is generated by any of its elements — this justifies the fact that the two optimization problems are equivalent.

More formally, the previous inequality constrained problem is equivalent to this equality constrained problem due to proposition 16 with $\{\mathbf{a} \in \mathbb{R}^p \mid \mathbf{a}^t \mathbf{W} \mathbf{a} \leq 1\}$ as closed convex bounded from below set. This is the way in which the optimization problem will be considered in this text.

As in the previous case, for fixed λ , the solutions \mathbf{a} can be interpreted geometrically in \mathbb{R}^p as the intersection of a one dimensional linear subspace and the surface determined by $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$, that is $V_{\mathbf{a}} \cap \{\mathbf{a}^t \mathbf{W} \mathbf{a} = 1\}$ (strictly, the point \mathbf{a} must be excluded).

CONVEXITY

In these problems, the objective function is convex due to proposition 13.d (in appendix C) and the positive definiteness of the matrix \mathbf{B} .

Remark 34 The search for a solution can be tackled via the previous equivalent optimization problems. In the first one, the feasible region is compact and, as $\lambda(\mathbf{a})$ is continuous, the Weierstrass theorem implies the existence of a solution.

Remark 35 The optimization problem B.17 can be seen as a particular case of *fractional optimization*, that is, the optimization of a quotient of functions. To see how these problems can be turned into nonfractional equality constrained problems, see proposition 27 (in appendix C).

B.2.2 Case $K = 2$: Two Populations

We have taken into consideration classification into two populations. It is well-known that this case can be written as an equivalent linear regression problem; nevertheless, we have not used this interpretation here.

On the other hand, since

$$(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}) = \frac{n_2}{n}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \quad (\text{B.25})$$

and

$$(\bar{\mathbf{x}}^{(2)} - \bar{\mathbf{x}}) = \frac{n_1}{n}(\bar{\mathbf{x}}^{(2)} - \bar{\mathbf{x}}^{(1)}) = -\frac{n_1}{n}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}), \quad (\text{B.26})$$

it follows that

$$\begin{aligned} \mathbf{a}^t \mathbf{B} \mathbf{a} &= \mathbf{a}^t \left[\sum_{k=1}^K n_k (\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(k)} - \bar{\mathbf{x}})^t \right] \mathbf{a} \\ &= \mathbf{a}^t \frac{n_1 n_2}{n} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{a} \\ &= \frac{n_1 n_2}{n} [\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})][\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^t \\ &= \frac{n_1 n_2}{n} [\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2. \end{aligned} \quad (\text{B.27})$$

Thus, the previous optimization problems are equivalent, respectively, to the following ones:

$$\mathbf{a} = \operatorname{argmax} \left\{ \frac{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t \mathbf{W} \mathbf{a}} \right\} \quad (\text{B.28})$$

and

$$\mathbf{a} = \operatorname{argmax} \{ [\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2 \} \quad \text{subject to} \quad \mathbf{a}^t \mathbf{W} \mathbf{a} = 1. \quad (\text{B.29})$$

Remark 36 Notice that with this formulation the numerator highlights the objective of the optimization problem: maximizing the difference between the means under control of the variability.

B.2.3 Interpretation of the Coefficients

Usually the variables of the vector \mathbf{x} have been measured using different scales: localization, variability or even units of measure. Then, the mathematical solution of the optimization problem provides values a_i not taking into account this fact. The function

$$y = a_1 x_1 + \dots + a_p x_p = \mathbf{a}^t \mathbf{x}, \quad (\text{B.30})$$

with $\mathbf{a} = (a_1, \dots, a_p)^t$, can, however, be used for classifying.

A possible transformation that can be applied to the previous values is a translation so that the origin of the axes coincides with the global centroid of the samples and a homothecy so that the coordinates refer to the standard deviation of each axis. Then,

$$\tilde{y} = b_0 + b_1x_1 + \dots + b_px_p = b_0 + \mathbf{b}^t\mathbf{x}, \quad (\text{B.31})$$

with $\mathbf{b} = (b_1, \dots, b_p)^t$, where b_i can be interpreted as regression coefficients. When these values are computed from crude data, they are termed *nonstandardized coefficients*, that represent the (absolut) contribution of the variables to the function but are not comparable among them. Anyway, \tilde{y} is also used for classifying. On the other hand, the typification of variables solves at the same time the above-mentioned scale problems —localization, variability and units—, so if \mathbf{b} is computed from typified —not crude— variables, this vector contains the *standardized coefficients*, that represent the relative contribution of the variables to the function and are comparable among them. Now, however, the function (B.31) cannot be used for classifying, as the important information has been lost (in this case $b_0 = 0$, for example).

Our proposal has been described in terms of the function (B.30) suggested by the optimization problem; nevertheless, the interpretation of the coefficients —through the figures— has been based on the function

$$y = \mathbf{a}^t\mathbf{x} = \mathbf{a}^t\mathbf{D}\mathbf{D}^{-1}\mathbf{x} = \mathbf{a}^t\mathbf{D}\tilde{\mathbf{x}}, \quad (\text{B.32})$$

where $\tilde{\mathbf{x}} = \mathbf{D}^{-1}\mathbf{x}$, with \mathbf{D} being the diagonal matrix with elements $\sigma_1, \dots, \sigma_p$, where σ_i is the standard deviation of the variable x_i . After applying this *univariate standardization*, the new variables have variance equal to one.

Note that the previous transformation does not change the mean of each variable. Thus, for the interpretation we have considered the sample version of the coefficients defined by $\mathbf{a}^t\mathbf{D}$, that is, the sample version of the quantities

$$\mathbf{a}^t\mathbf{D} = (a_1\sigma_1, \dots, a_p\sigma_p). \quad (\text{B.33})$$

B.3 The Discriminant Function

For two populations, the resolution of the optimization problem, with and without the classical constraint (when $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$ or $\beta = 0$, respectively), is given at the same time by

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \mathbf{a}} \left(\frac{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t \mathbf{W} \mathbf{a}} - \beta (\mathbf{a}^t \mathbf{W} \mathbf{a} - 1) \right) \\
&= \frac{\partial}{\partial \mathbf{a}} \left(\frac{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2}{\mathbf{a}^t \mathbf{W} \mathbf{a}} \right) - \beta 2 \mathbf{W} \mathbf{a} \\
&= \frac{2 \mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \mathbf{a}^t \mathbf{W} \mathbf{a} - [\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2 2 \mathbf{W} \mathbf{a}}{(\mathbf{a}^t \mathbf{W} \mathbf{a})^2} \\
&\quad - \beta 2 \mathbf{W} \mathbf{a}
\end{aligned} \tag{B.34}$$

so

$$\frac{\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \mathbf{a}^t \mathbf{W} \mathbf{a}}{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2 + \beta (\mathbf{a}^t \mathbf{W} \mathbf{a})^2} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) = \mathbf{W} \mathbf{a} \tag{B.35}$$

and, if \mathbf{W} is invertible,

$$\mathbf{a} = \frac{\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \mathbf{a}^t \mathbf{W} \mathbf{a}}{[\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})]^2 + \beta (\mathbf{a}^t \mathbf{W} \mathbf{a})^2} \mathbf{W}^{-1} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}). \tag{B.36}$$

Since $\mathbf{a}^t (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$ and $\mathbf{a}^t \mathbf{W} \mathbf{a}$ are numbers, it does not matter whether the constraint $\mathbf{a}^t \mathbf{W} \mathbf{a} = 1$ is imposed or not, the solution for the classical linear discriminant analysis is that y is proportional to $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}$, and, without loss of generality:

$$y = \mathbf{a}_F^t \mathbf{x} = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \mathbf{x}. \tag{B.37}$$

Since $y \in \mathbb{R}$, it is sometimes written as $y = y^t = \mathbf{x}^t \mathbf{a}_F = \mathbf{x}^t \mathbf{W}^{-1} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$ in the literature.

B.4 The Classification

Geometrically, the Fisher's discriminant analysis projects the data into one-dimensional linear subspaces (see addendum 1.2.4). For the first direction, this operation is analytically done by the $y = \mathbf{a}_F^t \mathbf{x}$ operation; that is, the multivariate vector \mathbf{x} is projected by the $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}$ premultiplication. The method classifies a new element in the population k as follows:

$$\begin{cases} k = 1 & \text{if } y > \frac{1}{2} (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} (\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\ k = 2 & \text{otherwise} \end{cases} \tag{B.38}$$

with the value $\frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$, the *cutoff point*, determined by the equation

$$(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) = 0, \quad (\text{B.39})$$

where the equality determines an hyperplane. By writing

$$\frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \quad (\text{B.40})$$

we see that the cutoff point is the projection of the midpoint between the two population sample averages, $\frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$, into the same subspace.

Thus, for $y = \mathbf{a}_F^t \mathbf{x}$ the classification of a multivariate point is made by the simple comparison of its projection with the projection of the semisum of the group means, taking into account that (see again section 1.2.4)

$$\|Proj_{V_{\mathbf{a}_F}}(\mathbf{x}) - Proj_{V_{\mathbf{a}_F}}(\bar{\mathbf{x}}^{(k)})\|_e = |\mathbf{a}_F^t \mathbf{x} - \mathbf{a}_F^t \bar{\mathbf{x}}^{(k)}|, \quad (\text{B.41})$$

where $\|\cdot\|_e$ denotes the Euclidean norm in \mathbb{R}^p and $V_{\mathbf{a}_F}$ is the linear subspace generated by \mathbf{a}_F .

Remark 37 Expressions (B.37) and (B.38) lead to the —also named in literature— *(sample) linear discriminant function*

$$\begin{aligned} L(\mathbf{x}) &= (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\ &= y(\mathbf{x}) - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \end{aligned} \quad (\text{B.42})$$

Notice that the second factor does not depend on the variables \mathbf{x} . See, for example, section 12.2.2 of Krzanowski (2000).

Remark 38 A comparison with the midpoint $\frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ is possible for two groups, but not for more populations. The following point of view of the classification rule will always apply. For each k , the centroid $\bar{\mathbf{x}}^{(k)}$ of each group can be interpreted as a representative multivariate point of the group, $\mathbf{r}^{(k)} = \bar{\mathbf{x}}^{(k)}$. Then, the previous classification criterion can be expressed as

$$k = \operatorname{argmin}_{\{1,2\}} \{d(y(\mathbf{x}), y(\mathbf{r}^{(k)}))\} = \operatorname{argmin}_{\{1,2\}} \{|\mathbf{a}_F^t \mathbf{x} - \mathbf{a}_F^t \mathbf{r}^{(k)}|\}. \quad (\text{B.43})$$

Equivalently, for the group k the quantity $y(\mathbf{r}^{(k)})$ can be thought of as a representative univariate point. It is important to note, however, that the classification with (B.38) does not take into account the magnitude —only the sign— of the distance to the midpoint, while the classification with (B.43) does.

Remark 39 The use of data and the previous optimization problem provide a value for \mathbf{a} . Then, if there is interest in the stochastic character of the vectors \mathbf{X} and Y , the following discriminant function

$$Y = \mathbf{a}_F^t \mathbf{X} = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1} \mathbf{X} \quad (\text{B.44})$$

and classification rule

$$\begin{cases} k = 1 & \text{if } Y > \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^t \mathbf{W}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\ k = 2 & \text{otherwise} \end{cases} \quad (\text{B.45})$$

can be considered instead of (B.37) and (B.38), where \mathbf{X} and Y are random variables again.

Remark 40 Finally, this appendix can be interpreted in the classification framework of section 1.5. Some population information is unknown and the use of samples allows inferring it; the sample rule (B.38) is obtained; on the other hand, an underlying theoretical rule can be obtained by substituting the corresponding population quantities into it.

Appendix C

Optimization Theory

All the contents of this section can be found in Bertsekas (1999). Let be $\mathcal{D} \subset \mathbb{R}^p$ and the differentiable function $f : \mathcal{D} \rightarrow \mathbb{R}$ depending on the variables $\mathbf{x} = (x_1, \dots, x_p)^t$, where $p \in \mathbb{N} \setminus \{0\}$ and \mathbb{R}^p is equipped with the norm $\|\cdot\| = \|\cdot\|_p$. Notice that since the function f is supposed to be differentiable, gradients and Taylor series can be used to study the behaviour and the conditions in the neighbourhood of the minima.

Some Definitions

MINIMA

Definition 44 *The point $\mathbf{x}_0 \in \mathcal{D}$ is said to be a (constrained) relative —or local— minimum if*

$$f(\mathbf{x}_0) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in D_\epsilon(\mathbf{x}_0), \quad (\text{C.1})$$

where $D_\epsilon(\mathbf{x}_0) = \{\mathbf{x} \mid \|\mathbf{x}_0 - \mathbf{x}\| < \epsilon\}$, for some $\epsilon > 0$ such that $D_\epsilon(\mathbf{x}_0) \subset \mathcal{D}$.

Definition 45 *The point $\mathbf{x}_0 \in \mathcal{D}$ is said to be a (constrained) global minimum if*

$$f(\mathbf{x}_0) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{D}. \quad (\text{C.2})$$

GENERAL PROBLEM

A general constrained optimization (minimization) problem with *feasible* or *admissible domain* \mathcal{D} consists of searching for the point \mathbf{x} such that

$$\mathbf{x} = \operatorname{argmin} \{f(\mathbf{x})\} \quad \text{subject to} \quad \mathbf{x} \in \mathcal{D}, \quad (\text{C.3})$$

where f is the *objective* —or *cost*— *function*. Usually, \mathcal{D} is expressed through some constraints on differentiable multivariate functions $\mathbf{h} = (h_1, \dots, h_{m_h})^t$ and $\mathbf{g} = (g_1, \dots, g_{m_g})^t$; for example, the *equality constraints* $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and the *inequality constraints* $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$, that would provide

$$\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \text{ and } \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}. \quad (\text{C.4})$$

The points of \mathcal{D} are termed *feasible* points.

EXISTENCE OF SOLUTIONS

It is well-known, due to the *Weierstrass theorem*, that there is at least one global minimum if f is a continuous function and \mathcal{D} is compact.

On the other hand, there is also an optimal solution when f is continuous, \mathcal{D} is closed and $f(\mathbf{x}_k) \rightarrow \infty$ for any $\{\mathbf{x}_k\}_k \subset \mathcal{D}$ such that $\|\mathbf{x}_k\| \rightarrow \infty$.

Some other results on the existence of optimal solutions are in section 2.1.2 of Bertsekas (1999) (and below in this appendix).

MAXIMIZATION PROBLEM

Local and global maximum are defined in the same way, just substituting \leq by \geq in (C.1) and (C.2).

The general minimization problem (C.3) is equivalent to the following maximization one:

$$\mathbf{x} = \operatorname{argmax} \{-f(\mathbf{x})\} \quad \text{subject to} \quad \mathbf{x} \in \mathcal{D}. \quad (\text{C.5})$$

In this case, all the forthcoming theory can be reformulated replacing the concepts related to minimum and convexity (of f) by the corresponding concepts of maximum and concavity (of f).

CONVEXITIES

Definition 46 *The subset $\mathcal{D} \subset \mathbb{R}^p$ is said to be convex if*

$$c\mathbf{x}_1 + (1 - c)\mathbf{x}_2 \in \mathcal{D} \quad (\text{C.6})$$

for all $\mathbf{x}_i \in \mathcal{D}$, $i = 1, 2$ and all $c \in [0, 1]$.

In the following, the domain \mathcal{D} is supposed to be convex, although some results also hold for an open set containing the minimum.

Definition 47 For a convex domain \mathcal{D} , a function $f : \mathcal{D} \rightarrow \mathbb{R}$ is convex if

$$f(c\mathbf{x}_1 + (1 - c)\mathbf{x}_2) \leq cf(\mathbf{x}_1) + (1 - c)f(\mathbf{x}_2) \quad (\text{C.7})$$

for all $\mathbf{x}_i \in \mathcal{D}$, $i = 1, 2$ and all $c \in [0, 1]$, and f is strictly convex if the previous inequality holds with $<$ instead of \leq , for all $\mathbf{x}_i \in \mathcal{D}$, $i = 1, 2$ with $\mathbf{x}_1 \neq \mathbf{x}_2$ and all $c \in (0, 1)$.

Proposition 13 Let $\mathcal{D} \subset \mathbb{R}^p$ be a convex set and let $f : \mathcal{D} \rightarrow \mathbb{R}$ be twice continuously differentiable, then

- (a) If $\nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in \mathbb{R}^p$, then f is convex over \mathcal{D} .
- (b) If $\nabla^2 f(\mathbf{x})$ is positive definite for every $\mathbf{x} \in \mathbb{R}^p$, then f is strictly convex over \mathcal{D} .
- (c) If $\mathcal{D} = \mathbb{R}^p$ and f is convex, then $\nabla^2 f(\mathbf{x})$ is positive semidefinite for all $\mathbf{x} \in \mathcal{D}$.
- (d) The quadratic function $f(\mathbf{x}) = \mathbf{x}^t \mathbf{M} \mathbf{x}$, where \mathbf{M} is a symmetric $p \times p$ matrix, is convex if and only if \mathbf{M} is positive semidefinite. Furthermore, f is strictly convex if and only if \mathbf{M} is positive definite.

Proof. See appendix B.1 of Bertsekas (1999). □

Proposition 14

- (a) For any collection $\{\mathcal{D}_i \mid i \in I\}$ of convex sets, the set intersection $\bigcap_{i \in I} \mathcal{D}_i$ is convex.
- (b) The image of a convex set under a linear transformation is convex.
- (c) If \mathcal{D} is a convex set and $f : \mathcal{D} \rightarrow \mathbb{R}$ is a convex function, the level sets $\{\mathbf{x} \in \mathcal{D} \mid f(\mathbf{x}) \leq c\}$ and $\{\mathbf{x} \in \mathcal{D} \mid f(\mathbf{x}) < c\}$ are convex for all scalars c .

Proof. See appendix B.1 of Bertsekas (1999). □

Convexities and Optimization

Taking into account that an unconstrained problem is a particular case of the constrained one, with $\mathcal{D} = \mathbb{R}^p$, the following statements concern both situations.

Definition 48 An optimization problem is convex if f and \mathcal{D} are convex.

The importance of this definition is that, in a convex problem, the local minimum —when it exists— is the global minimum.

Proposition 15 If f is a convex function and \mathcal{D} is also convex, then a local minimum of f over \mathcal{D} is a global minimum. If in addition f is strictly convex over \mathcal{D} , then at most one global minimum of f over \mathcal{D} exists.

Proof. See section 2.1 of Bertsekas (1999). □

Definition 49 The point $\mathbf{x}_0 \in \mathcal{D} \subset \mathbb{R}^p$ is said to be an extreme —or frontier— point if for any neighbourhood $N(\mathbf{x}_0)$ such that $\mathbf{x}_0 \in N(\mathbf{x}_0) \subset \mathcal{D}$ there are both some point belonging to \mathcal{D} and some point not belonging to \mathcal{D} .

Proposition 16 Let f be convex, let \mathcal{D} be closed convex bounded from below, then if f attains a maximum over \mathcal{D} , it attains a maximum at some extreme point of \mathcal{D} .

Proof. See appendix B.4 of Bertsekas (1999). □

As in the unconstrained optimization, the first order variation $\nabla f(\mathbf{x}_0)^t \Delta \mathbf{x}$, due to a small feasible variation $\Delta \mathbf{x}$, is expected to be nonnegative at a local minimum \mathbf{x}_0 . Due to the convexity of \mathcal{D} , the feasible variations are of the form $\Delta \mathbf{x} = \mathbf{x} - \mathbf{x}_0$, with $\mathbf{x} \in \mathcal{D}$.

Proposition 17 (*First Order Necessary Condition*)

(a) If \mathbf{x}_0 is a local minimum of f over \mathcal{D} , then

$$\nabla f(\mathbf{x}_0)^t (\mathbf{x} - \mathbf{x}_0) \geq 0, \quad \forall \mathbf{x} \in \mathcal{D}. \quad (\text{C.8})$$

(b) If f is convex over \mathcal{D} , then the condition of part (a) is also sufficient for \mathbf{x}_0 to minimize f over \mathcal{D} .

Proof. See section 2.1 of Bertsekas (1999). □

Remark 41 In the unconstrained situation, when f is convex the first order condition $\nabla f(\mathbf{x}_0) = 0$ (given in C.18) is also sufficient for optimality.

Proposition 18 (*Second Order Necessary Condition*) If \mathbf{x}_0 is a local minimum of the twice continuously differentiable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ over the convex set \mathcal{D} , then

$$(\mathbf{x} - \mathbf{x}_0)^t \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) \geq 0 \quad (\text{C.9})$$

for all $\mathbf{x} \in \mathcal{D}$ such that $\nabla f(\mathbf{x}_0)^t (\mathbf{x} - \mathbf{x}_0) = 0$.

Proof. See exercise 2.1.10 of Bertsekas (1999). □

With respect to the sufficient conditions, the following result holds.

Proposition 19 (*Second Order Sufficient Condition*) If \mathbf{x}_0 is a local minimum of the twice continuously differentiable function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ over the convex set \mathcal{D} , then

$$\nabla f(\mathbf{x}_0)^t (\mathbf{x} - \mathbf{x}_0) \geq 0 \quad \forall \mathbf{x} \in \mathcal{D} \quad (\text{C.10})$$

and one of the following three conditions holds:

1. \mathcal{D} is polyhedral and it follows that

$$(\mathbf{x} - \mathbf{x}_0)^t \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) > 0 \quad (\text{C.11})$$

for all $\mathbf{x} \in \mathcal{D}$ satisfying $\mathbf{x} \neq \mathbf{x}_0$ and $\nabla f(\mathbf{x}_0)^t (\mathbf{x} - \mathbf{x}_0) = 0$.

2. It follows that $\bar{\mathbf{x}}^t \nabla^2 f(\mathbf{x}_0) \bar{\mathbf{x}} > 0$ for all nonzero $\bar{\mathbf{x}}$ that are in the closure of the set $\{d \mid d = \alpha(\mathbf{x} - \mathbf{x}_0), \mathbf{x} \in \mathcal{D}, \alpha \geq 0\}$ and satisfy $\nabla f(\mathbf{x}_0)^t \bar{\mathbf{x}} = 0$.

3. For some $\gamma > 0$, we have

$$(\mathbf{x} - \mathbf{x}_0)^t \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) > \gamma \|\mathbf{x} - \mathbf{x}_0\|^2, \quad \mathbf{x} \in \mathcal{D}. \quad (\text{C.12})$$

Proof. See exercise 2.1.11 of Bertsekas (1999). □

It is worthwhile noting how, due to the monotony of a convex function and the topology of a convex domain, the optimization problem is simplified in two directions:

1. **Minimum.** A local minimum is also a global minimum: see proposition 15.
2. **Maximum.** There is a extreme (or frontier) point of the domain where the maximum —when exists— is attained — see proposition 16.

Unconstrained Problem

The unconstrained situation is a particular case of the constrained situation with $\mathcal{D} = \mathbb{R}^p$. Nevertheless, this particular case is of great importance, as one of the approaches towards solving the constrained optimization consists in tackling an equivalent unconstrained problem. Minima are now defined as

Definition 50 *The point $\mathbf{x}_0 \in \mathbb{R}^p$ is said to be an unconstrained relative —or local— minimum if*

$$f(\mathbf{x}_0) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in D_\epsilon(\mathbf{x}_0), \quad (\text{C.13})$$

where $D_\epsilon(\mathbf{x}_0) = \{\mathbf{x} \mid \|\mathbf{x}_0 - \mathbf{x}\| < \epsilon\}$, for some $\epsilon > 0$.

Definition 51 *The point $\mathbf{x}_0 \in \mathbb{R}^p$ is said to be a unconstrained global minimum if*

$$f(\mathbf{x}_0) \leq f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p. \quad (\text{C.14})$$

NECESSARY CONDITIONS

Small variations $\Delta \mathbf{x}$ from \mathbf{x}_0 yield, up to first order, a cost variation

$$f(\mathbf{x}_0 + \Delta \mathbf{x}) - f(\mathbf{x}_0) \approx \nabla f(\mathbf{x}_0)^t \Delta \mathbf{x} \quad (\text{C.15})$$

and, up to second order,

$$f(\mathbf{x}_0 + \Delta \mathbf{x}) - f(\mathbf{x}_0) \approx \nabla f(\mathbf{x}_0)^t \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^t \nabla^2 f(\mathbf{x}_0) \Delta \mathbf{x}. \quad (\text{C.16})$$

It is expected that if \mathbf{x}_0 is an unconstrained local minimum, the first order cost variation due to small variation $\Delta \mathbf{x}$ is nonnegative

$$\nabla f(\mathbf{x}_0)^t \Delta \mathbf{x} = \sum_{i=1}^p \frac{\partial f(\mathbf{x}_0)}{\partial x_i} \Delta x_i \geq 0, \quad (\text{C.17})$$

and, in particular, by taking $\Delta \mathbf{x}$ to be positive and negative multiples of the unit coordinate vectors, we obtain $\frac{\partial f(\mathbf{x}_0)}{\partial x_i} \geq 0$ and $\frac{\partial f(\mathbf{x}_0)}{\partial x_i} \leq 0$, respectively, so the equivalent necessary condition

$$\nabla f(\mathbf{x}_0) = 0 \quad (\text{C.18})$$

is obtained.

On the other hand, it is also expected that the second order cost variation due to small $\Delta \mathbf{x}$ must be nonnegative

$$\nabla f(\mathbf{x}_0)^t \Delta \mathbf{x} + \frac{1}{2} \Delta \mathbf{x}^t \nabla^2 f(\mathbf{x}_0) \Delta \mathbf{x} \geq 0,$$

that becomes (by applying C.18)

$$\Delta \mathbf{x}^t \nabla^2 f(\mathbf{x}_0) \Delta \mathbf{x} \geq 0, \quad (\text{C.19})$$

which implies that the matrix

$$\nabla^2 f(\mathbf{x}_0) \quad (\text{C.20})$$

is positive semidefinite.

Proposition 20 *Let \mathbf{x}_0 be an unconstrained local minimum of $f : \mathbb{R}^p \rightarrow \mathbb{R}$ and assume that f is continuously differentiable in an open set \mathcal{D} containing \mathbf{x}_0 ; then*

$$\nabla f(\mathbf{x}_0) = 0 \quad (\text{C.21})$$

and if, in addition, f is twice continuously differentiable within \mathcal{D} , the

$$\nabla^2 f(\mathbf{x}_0) \text{ is positive semidefinite.} \quad (\text{C.22})$$

Proof. See section 1.1 of Bertsekas (1999). □

SUFFICIENT CONDITIONS

If \mathbf{x}_0 verifies

$$\nabla f(\mathbf{x}_0) = 0 \quad (\text{C.23})$$

and

$$\nabla^2 f(\mathbf{x}_0) \text{ is positive definite} \quad (\text{C.24})$$

then for all $\Delta \mathbf{x} \neq \mathbf{0}$ it holds that

$$\Delta \mathbf{x}^t \nabla^2 f(\mathbf{x}_0) \Delta \mathbf{x} > 0, \quad (\text{C.25})$$

implying that at \mathbf{x}_0 the second order variation of f due to small nonzero variation $\Delta \mathbf{x}$ is positive. Thus, the function f tends to increase strictly with small departures from \mathbf{x}_0 , suggesting that the conditions (C.23) and (C.24) are sufficient for local optimality.

Proposition 21 *Let $f : \mathbb{R}^p \rightarrow \mathbb{R}$ be twice continuously differentiable in an open set \mathcal{D} ; suppose that $\mathbf{x}_0 \in \mathcal{D}$ satisfies the conditions*

$$\nabla f(\mathbf{x}_0) = 0 \quad (\text{C.26})$$

and

$$\nabla^2 f(\mathbf{x}_0) \text{ is positive definite (not only semidefinite);} \quad (\text{C.27})$$

then \mathbf{x}_0 is a strict unconstrained local minimum of f . In particular, scalars $\gamma > 0$ and $\epsilon > 0$ exist such that

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_0\|^2 \quad \forall \mathbf{x} \text{ such that } \|\mathbf{x} - \mathbf{x}_0\| < \epsilon. \quad (\text{C.28})$$

Proof.

See section 1.1 of Bertsekas (1999). □

Constrained Problem: Equality Constraints

Under some conditions, the *constrained* optimization problem (C.3) can be converted in an equivalent *unconstrained* optimization problem. Let

$$\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{h}(\mathbf{x}) = \mathbf{0}\}. \quad (\text{C.29})$$

NECESSARY CONDITIONS

Proposition 22 (Lagrange Multiplier Theorem) *Let \mathbf{x}_0 be the local minimum of f subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, and assume that the vectors $\nabla h_i(\mathbf{x}_0)$, $i = 1, \dots, m_h$ are linearly independent. Then there is a unique (column) vector β_0 , called a Lagrange multiplier vector, such that*

$$\nabla f(\mathbf{x}_0) + \sum_{i=1}^{m_h} \beta_i \nabla h_i(\mathbf{x}_0) = \mathbf{0}. \quad (\text{C.30})$$

If in addition f and \mathbf{h} are twice continuously differentiable, we have

$$\mathbf{x}^t \left(\nabla^2 f(\mathbf{x}_0) + \sum_{i=1}^{m_h} \beta_i \nabla^2 h_i(\mathbf{x}_0) \right) \mathbf{x} \geq \mathbf{0} \quad (\text{C.31})$$

for all $\mathbf{x} \in V(\mathbf{x}_0) = \{\mathbf{x} \mid \nabla h_i(\mathbf{x}_0)^t \mathbf{x} = 0, i = 1, \dots, m_h\}$.

Proof.

See section 3.1 of Bertsekas (1999).

□

Both local minima and local maxima —and possible other points— may satisfy the first order necessary conditions. In this situation, the second order necessary conditions are used to find local minima.

THE LAGRANGIAN

It is useful to combine the objective function and the constraints in the following function.

Definition 52 *The Lagrangian of the optimization problem (C.3) with domain (C.29) is given by*

$$F(\mathbf{x}, \beta) = f(\mathbf{x}) + \beta^t \mathbf{h}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{D}. \quad (\text{C.32})$$

where $\beta = (\beta_1, \dots, \beta_{m_h})^t$ are the multipliers.

SUFFICIENT CONDITIONS

Proposition 23 (Second Order Conditions) *Let f and \mathbf{h} be twice continuously differentiable, and let $\mathbf{x}_0 \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}^{m_h}$ such that*

$$\nabla_{\mathbf{x}} F(\mathbf{x}_0, \beta_0) = \mathbf{0} \quad \text{and} \quad \nabla_{\beta} F(\mathbf{x}_0, \beta_0) = \mathbf{0} \quad (\text{C.33})$$

and

$$\mathbf{x}^t \nabla_{\mathbf{xx}}^2 F(\mathbf{x}_0, \beta_0) \mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0} \quad \text{with} \quad \nabla \mathbf{h}(\mathbf{x}_0)^t \mathbf{x} = 0. \quad (\text{C.34})$$

Then \mathbf{x}_0 is a strict local minimum of f subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$. In fact, scalars $\gamma > 0$ and $\epsilon > 0$ exist such that

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \frac{\gamma}{2} \|\mathbf{x} - \mathbf{x}_0\|^2 \quad \forall \mathbf{x} \quad \text{such that} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \quad \text{and} \quad \|\mathbf{x} - \mathbf{x}_0\| < \epsilon.$$

Proof.

See section 3.2 of Bertsekas (1999).

□

EQUIVALENT UNCONSTRAINED PROBLEM

In short, the constrained optimization problem can be studied via the following unconstrained optimization problem

$$\mathbf{x} = \operatorname{argmin} \{F(\mathbf{x}, \beta)\}. \quad (\text{C.35})$$

Constrained Problem: Inequality Constraints

Let the feasible domain given by

$$\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{h}(\mathbf{x}) = \mathbf{0} \text{ and } \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}. \quad (\text{C.36})$$

THE LAGRANGIAN

In this situation, the Lagrangian takes the form

Definition 53 *The Lagrangian of the optimization problem (C.3) with domain (C.36) is given by*

$$F(\mathbf{x}, \beta, \mu) = f(\mathbf{x}) + \beta^t \mathbf{h}(\mathbf{x}) + \mu^t \mathbf{g}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{D}. \quad (\text{C.37})$$

where $\beta = (\beta_1, \dots, \beta_{m_g})^t$ and $\mu = (\mu_1, \dots, \mu_{m_h})^t$ are the multipliers.

Remark 42 The sign of the added terms is usually taken as positive; for example, the last term is usually taken as $+\mu^t \mathbf{g}(\mathbf{x})$, instead of $-\mu^t \mathbf{g}(\mathbf{x})$; nevertheless, we have used the latter form in chapter three due to the particular function $\mathbf{g}(\mathbf{x}) = -\mathbf{x}$.

There are several ways of dealing with the inequality constraints problem, one of them is based on the previous equality constraints framework. The following approach is not the most general, since some regularity conditions are needed, but it is the most direct generalization of the previous theory (notice that the generalised Rayleigh quotient $\lambda(\mathbf{a})$ fulfils the regularity conditions).

NECESSARY CONDITIONS

The following proposition generalises the Lagrange multiplier theorem (proposition 22).

Proposition 24 (*Karush-Kuhn-Tucker Conditions*) *Let \mathbf{x}_0 be the local minimum of f subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$, where f , \mathbf{h} and \mathbf{g} are continuously differentiable functions, and assume that $\nabla h_i(\mathbf{x}_0)$, $i = 1, \dots, m_h$ and $\nabla g_i(\mathbf{x}_0)$, $i = 1, \dots, m_g$ are linearly independent; then unique Lagrange multiplier vectors β_0 and μ_0 exists such that*

$$\nabla_{\mathbf{x}} F(\mathbf{x}_0, \beta_0, \mu_0) = \mathbf{0} \quad (\text{C.38})$$

$$\mu \geq \mathbf{0} \quad \text{with} \quad \mu_i = 0 \quad \text{when} \quad g_i(\mathbf{x}_0) < 0. \quad (\text{C.39})$$

If in addition f , \mathbf{h} and \mathbf{g} are twice continuously differentiable, there holds

$$\mathbf{x}^t \nabla_{\mathbf{xx}}^2 F(\mathbf{x}_0, \beta_0, \mu_0) \mathbf{x} \geq \mathbf{0} \quad (\text{C.40})$$

for all \mathbf{x} such that $\nabla h_i(\mathbf{x}_0)^t \mathbf{x} = 0$, $i = 1, \dots, m_h$, and $\nabla g_i(\mathbf{x}_0)^t \mathbf{x} = 0$ when $g_i(\mathbf{x}_0) = 0$.

Proof. See section 3.3 of Bertsekas (1999). □

SEARCHING STRATEGY

As Bertsekas (1999) states: *One approach for using necessary conditions to solve inequality constrained problems is to consider separately all the possible combinations of constraints being active or inactive* (an inequality constraint $g_i(\mathbf{x}) \leq 0$ is active at point \mathbf{x} if the equality holds, and inactive otherwise).

SUFFICIENT CONDITIONS

Proposition 25 (Second Order Conditions) *Let f , \mathbf{h} and \mathbf{g} be twice continuously differentiable, and let $\mathbf{x}_0 \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^{m_h}$ and $\mu_0 \in \mathbb{R}^{m_g}$ such that*

$$\nabla_{\mathbf{x}} F(\mathbf{x}_0, \beta_0, \mu_0) = \mathbf{0}, \quad \mathbf{h}(\mathbf{x}_0) = \mathbf{0}, \quad \mathbf{g}(\mathbf{x}_0) \leq \mathbf{0} \quad (\text{C.41})$$

$$\mu \geq \mathbf{0} \quad \text{with} \quad \mu_i = 0 \quad \text{when} \quad g_i(\mathbf{x}_0) < 0. \quad (\text{C.42})$$

$$\mathbf{x}^t \nabla_{\mathbf{xx}}^2 F(\mathbf{x}_0, \beta_0, \mu_0) \mathbf{x} > \mathbf{0} \quad (\text{C.43})$$

for all \mathbf{x} such that $\nabla h_i(\mathbf{x}_0)^t \mathbf{x} = 0$, $i = 1, \dots, m_h$, and $\nabla g_i(\mathbf{x}_0)^t \mathbf{x} = 0$ when $g_i(\mathbf{x}_0) = 0$. Assume also that

$$\mu_i > 0 \quad \text{when} \quad g_i(\mathbf{x}_0) = 0. \quad (\text{C.44})$$

Then \mathbf{x}_0 is a strict local minimum of f subject to $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$.

Proof. See section 3.3 of Bertsekas (1999). □

This result is based on the transformation of the problem to one with equality constraints, and there are stricter versions of sufficient conditions.

Remark 43 Again from Bertsekas (1999): *The sufficient conditions that we have discussed so far [previous proposition] involve second derivatives and Hessian positive definiteness assumptions.*

Our experience with unconstrained problems suggests that the first order Lagrange multiplier conditions together with convexity assumptions should also be sufficient for optimality. Indeed this is so, as we will demonstrate shortly. In fact we will not need to impose convexity or even differentiability assumptions explicitly. A minimization condition on the Lagrangian function turns out to be sufficient.

Proposition 26 (General Conditions) *Let the problem be*

$$\mathbf{x} = \operatorname{argmin} \{f(\mathbf{x})\} \quad \text{subject to} \quad \mathbf{x} \in \mathcal{S} \quad \text{and} \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \quad (\text{C.45})$$

where f and \mathbf{g} are real valued functions on \mathbb{R}^p and \mathcal{S} is a given subset of \mathbb{R}^p . Let \mathbf{x}_0 be a feasible point which together with $\mu_0 \in \mathbb{R}^{m_g}$ satisfies

$$\mu \geq \mathbf{0} \quad \text{with} \quad \mu_i = 0 \quad \text{when} \quad g_i(\mathbf{x}_0) < 0 \quad (\text{C.46})$$

and minimizes the Lagrangian function $F(\mathbf{x}, \mu_0)$ over $\mathbf{x} \in (S)$, that is:

$$\mathbf{x}_0 = \operatorname{argmin} \{F(\mathbf{x}, \mu_0)\}; \quad (\text{C.47})$$

then \mathbf{x}_0 is a global minimum of the problem.

Proof. See section 3.3 of Bertsekas (1999). □

Remark 44 Note that the function f of this proposition can be a Lagrangian taking implicitly into account equality constraints.

EQUIVALENT UNCONSTRAINED PROBLEM

In short, the constrained —in \mathbf{x} — optimization problem can be studied through the following unconstrained —in \mathbf{x} — optimization problem

$$\mathbf{x} = \operatorname{argmin} \{F(\mathbf{x}, \beta, \mu)\} \quad \text{subject to} \quad \mu \geq \mathbf{0} \quad \text{with} \quad \mu_i = 0 \quad \text{when} \quad g_i(\mathbf{x}) < 0. \quad (\text{C.48})$$

KARUSH-KUHN-TUCKER CONDITIONS

From the previous propositions, it is clear that the following definition is quite useful for referring to the necessary and sufficient conditions.

The *Karush-Kuhn-Tucker conditions* (of first order) are given by

$$\begin{cases} \nabla_{\mathbf{x}} F = \mathbf{0} \\ \nabla_{\beta} F = \mathbf{0} \\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mu \geq \mathbf{0} \text{ and } \mu^t \mathbf{g}(\mathbf{x}) = \mathbf{0} \end{cases} \quad (\text{C.49})$$

or, in the usual differentiation notation, by

$$\begin{cases} \frac{\partial F}{\partial \mathbf{x}} = \mathbf{0} \\ \frac{\partial F}{\partial \beta} = \mathbf{0} \\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \mu \geq \mathbf{0} \text{ and } \mu^t \mathbf{g}(\mathbf{x}) = \mathbf{0} \end{cases} . \quad (\text{C.50})$$

Note that the two first are related to minimizing the Lagrangian.

NONNEGATIVITY CONSTRAINTS

For the particular case $\mathbf{g}(\mathbf{x}) = -\mathbf{x}$, that is, when the inequality constraints are $\mathbf{x} \geq \mathbf{0}$, the conditions are given by

$$\begin{cases} \frac{\partial F}{\partial \mathbf{x}} = \mathbf{0} \\ \frac{\partial F}{\partial \beta} = \mathbf{0} \\ x_i \geq 0, \mu_i \geq 0 \text{ and } \mu_i x_i = 0. \end{cases} \quad (\text{C.51})$$

One Particular Problem

In this section, some problems similar to those of our framework are included.

FRACTIONAL PROGRAMMING

Of special interest for our optimization problem is the following situation. Consider

Proposition 27

$$\mathbf{x} = \operatorname{argmin} \left\{ \frac{f(\mathbf{x})}{g(\mathbf{x})} \right\} \quad \text{subject to} \quad \mathbf{x} \in \mathcal{D}, \quad (\text{C.52})$$

where f and g are real functions on \mathbb{R}^p and \mathcal{D} is a given subset such that $g(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{D}$.

For $\beta \in \mathbb{R}$, it can be defined

$$Q(\beta) = \min_{\mathbf{x} \in \mathcal{D}} \{f(\mathbf{x}) - \beta g(\mathbf{x})\}. \quad (\text{C.53})$$

Suppose that a scalar β_0 and a vector \mathbf{x}_0 satisfy $Q(\beta_0) = 0$ and

$$\mathbf{x}_0 = \operatorname{argmin}_{\mathbf{x} \in \mathcal{D}} \{f(\mathbf{x}) - \beta_0 g(\mathbf{x})\}, \quad (\text{C.54})$$

respectively; then \mathbf{x}_0 is an optimal solution to the original problem.

Proof. See exercise 2.1.17 of Bertsekas (1999). □

Remark 45 Using this fact, it is possible to reduce the problem to an equivalent problem without the fraction of functions.

References

- [1] Abraham, C., P.A. Cornillon, E. Matzner-Løber and N. Molinari (2003). Unsupervised Curve Clustering Using B-Splines. *Scandinavian Journal of Statistics*. 30 (3), 581–595.
- [2] Abraham, C., G. Biau and B. Cadre (2006). On the Kernel Rule for Function Classification. *AIISM*. 58, 619–633.
- [3] Baïllo, A., and A. Cuevas (2008). Supervised Functional Classification: A Theoretical Remark and Some Comparisons. *Manuscript available at* <http://arxiv.org/abs/0806.2831>.
- [4] Berlinet, A., G. Biau and L. Rouvière (2008). Functional Supervised Classification with Wavelets. *Annales de l'I.S.U.P.*. 52 (1–2), 61–80.
- [5] Bertsekas, D.P. (1999). *Nonlinear Programming*. Athena Scientific (Second edition. Second printing 2003).
- [6] Biau, G., F. Bunea and M.H. Wegkamp (2003). Functional Classification in Hilbert Spaces. *IEEE Transactions on Information Theory*. 1 (11), 1–8.
- [7] Blandford, R.R. (1993). Discrimination of Earthquakes and Explosions at Regional Distances Using Complexity. *Report AFTAC-TR-93-044 HQ*, Air Force Technical Applications Center, Patrick Air Force Base, FL.
- [8] Boor, C. de (1978). *A Practical Guide to Splines*. Springer-Verlag.
- [9] Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press (First edition. Sixth printing 2008). *Book available at* http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.
- [10] Caiado, J., N. Crato and D. Peña (2006). A Periodogram-Based Metric for Time Series Classification. *Computational Statistics & Data Analysis*. 50, 2668–2684.

- [11] Chandler, G., and W. Polonik (2006). Discrimination of Locally Stationary Time Series Based on the Excess Mass Functional. *Journal of the American Statistical Association*. 101 (473), 240–253.
- [12] Cuesta-Albertos, J.A., and R. Fraiman (2007). Impartial Trimmed k -Means for Functional Data. *Computational Statistics & Data Analysis*. 51, 4864–4877.
- [13] Cuevas, A., M. Febrero and R. Fraiman (2007). Robust Estimation and Classification for Functional Data via Projection-Based Depth Notions. *Computational Statistics*. 22, 481–496.
- [14] Dabo-Niang, S., F. Ferraty and P. Vieu (2006). Mode Estimation for Functional Random Variable and Its Application for Curves Classification. *Far East Journal of Theoretical Statistics*. 18 (1), 93–119.
- [15] Dabo-Niang, S., F. Ferraty and P. Vieu (2007). On the Using of Modal Curves for Radar Waveforms Classification. *Computational Statistics & Data Analysis*. 51, 4878–4890.
- [16] Dahlhaus, R. (1996). Asymptotic Statistical Inference for Nonstationary Processes with Evolutionary Spectra. *Athens Conference on Applied Probability and Time Series Analysis, Vol. 2* (P.M. Robinson and M. Rosenblatt, eds.). *Lecture Notes in Statist.* 115 145–159. Springer-Verlag. 6, 171–191.
- [17] Dahlhaus, R., and Polonik, W. (2006). Nonparametric Quasi-Maximum Likelihood Estimation for Gaussian Locally Stationary Processes. *The Annals of Statistics*. 34 (6), 2790–2824.
- [18] Dahlhaus, R., and Polonik, W. (2009). Empirical Spectral Processes for Locally Stationary Time Series. *Bernoulli*. 15 (1), 1–39.
- [19] Diggle, P.J., (1990). *Time Series: A Biostatistical Introduction*. Clarendon Press.
- [20] Ferraty, F., and P. Vieu (2003). Curves Discrimination: A Nonparametric Functional Approach. *Computational Statistics & Data Analysis*. 44, 161–173.
- [21] Ferraty, F., and P. Vieu (2006). *Nonparametric Functional Data Analysis*. Springer.
- [22] Fisher, R.A.(1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*. 7 (2), 179–188.

- [23] Fraiman, R., and G. Muniz (2001). Trimmed Means for Functional Data. *Test*. 10 (2), 419-440.
- [24] Grimmett, G., and D. Stirzaker (2001). *Probability and Random Processes*. Oxford University Press (Third edition).
- [25] Hall, P., D.S. Poskitt and B. Presnell (2001). A Functional Data-Analytic Approach to Signal Discrimination. *Technometrics*. 43 (1), 1-9.
- [26] Hastie, T., A. Buja and R.J. Tibshirani (1995). Penalized Discriminant Analysis. *The Annals of Statistics*. 23 (1), 73-102.
- [27] Hirukawa, J. (2004). Discriminant Analysis for Multivariate Non-Gaussian Locally Stationary Processes. *Scientiae Mathematicae Japonicae*. 60 (2), 357-380.
- [28] Huang, H., H. Ombao and D.S. Stoffer (2004). Discrimination and Classification of Nonstationary Time Series Using the SLEX Model. *Journal of the American Statistical Association*. 99 (467), 763-774.
- [29] Jagannathan, R., and T. Ma (2003). Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraint Helps. *The Journal of Finance*. 58 (4), 1651-1683.
- [30] James, G.M., and T. Hastie (2001). Functional Linear Discriminant Analysis for Irregularly Sampled Curves. *Journal of the Royal Statistical Society. Series B*, 63, 533-550.
- [31] James, G.M., and C. A. Sugar (2003). Clustering for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*. 98 (462), 397-408.
- [32] Kakizawa, Y., R.H. Shumway and M. Taniguchi (1998). Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association*. 93 (441), 328-340.
- [33] Kiers, H.A.L. (1995). Maximization of Sums of Quotients of Quadratic Forms and Some Generalizations. *Psychometrika*. 60 (2), 221-245.
- [34] Krzanowski, W.J. (2000). *Principles of Multivariate Analysis*. Oxford University Press (First published 1988. Revised edition 2000).
- [35] Li, B., and Q. Yu (2008). Classification of Functional Data: A Segmentation Approach. *Computational Statistics & Data Analysis*. 52, 4790-4800.

- [36] López-Pintado, S., and J. Romo (2006). Depth-Based Classification for Functional Data. In *Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications*. American Mathematical Society. DIMACS Series, 72, 103–121. (R. Liu, R. Serfling and D.L. Souvaine [eds]).
- [37] López-Pintado, S., and J. Romo (2009). On the Concept of Depth for Functional Data. *Journal of the American Statistical Association*. 104 (486), 718–734.
- [38] Lütkepohl, H. (1996). *Handbook of Matrices*. John Wiley & Sons.
- [39] Maharaj, E.A., and A.M. Alonso (2007). Discrimination of Locally Stationary Time Series Using Wavelets. *Computational Statistics & Data Analysis*. 52, 879–895.
- [40] Mardia, K.V., J.T. Kent and J.M. Bibby (1979). *Multivariate Analysis*. Academic Press.
- [41] Marsden, J.E., and A.J. Tromba (1991). *Cálculo vectorial*. Addison-Wesley Iberoamericana.
- [42] McDonald, R.P. (1979). Some Results on Proper Eigenvalues and Eigenvectors with Applications to Scaling. *Psychometrika*. 44 (2), 211–227.
- [43] Nerini, D., and B. Ghattas (2007). Classifying Densities Using Functional Regression Trees: Applications in Oceanology. *Computational Statistics & Data Analysis*. 51, 4984–4993.
- [44] Nualart, D. (1995). *The Malliavin Calculus and Related topics*. Springer-Verlag.
- [45] Ombao, H.C., J.A. Raz, R. von Sachs and B.A. Malow (2001). Automatic Statistical Analysis of Bivariate Nonstationary Time Series. *Journal of the American Statistical Association*. 96 (454), 543–560.
- [46] Priestley, M. (1965). Evolutionary Spectra and Non-stationary Processes. *Journal of the Royal Statistical Society. Series B*, 27 (2), 204–237.
- [47] Priestley, M.B. (1981). *Spectral Analysis and Time Series*. Elsevier (Eleventh printing 2001. Reprinted 2004).
- [48] Ramsay, J.O., and B.W. Silverman (2006). *Functional Data Analysis*. Springer.
- [49] Rossi, F., and N. Villa (2006). Support Vector Machine for Functional Data Classification. *Neurocomputing*. 69, 730–742.

- [50] Sakiyama, K., and M. Taniguchi (2004). Discriminant Analysis for Locally Stationary Processes. *Journal of Multivariate Analysis*. 90, 282–300.
- [51] Shumway, R.S. (2003). Time-Frequency Clustering and Discriminant Analysis. *Statistics & Probability Letters*. 63, 307–314.
- [52] Shumway, R.H., and D.S. Stoffer (2000). *Time Series Analysis and Its Applications*. Springer.
- [53] Taniguchi, M., and Y. Kakizawa (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer.
- [54] Tarpey, T., and K.K.J. Kinader (2003). Clustering Functional Data. *Journal of Classification*. 20 (1), 93–114.
- [55] Vardi, Y., and C.H. Zhang (2000). The Multivariate L_1 -Median and Associated Data Depth. *Proceedings of the National Academy of Science USA*. 97 (4), 1423–1426.
- [56] Wang, K., and T. Gasser (1997). Alignment of Curves by Dynamic Time Warping. *The Annals of Statistics*. 25 (3), 1251–1276.
- [57] Wang, K., and T. Gasser (1999). Synchronizing Sample Curves Non-Parametrically. *The Annals of Statistics*. 27 (2), 439–460.
- [58] Wold, H.O.A. (1938). *A Study in the Analysis of Stationary Time Series*. Almqvist and Wiksell.
- [59] Zuo, Y. (2003). Projected-Based Depth Functions and Associated Medians. *The Annals of Statistics*. 31 (5), 1460–1490.
- [60] Zuo, Y., and R. Serfling (2000). General Notions of Statistical Depth Function. *The Annals of Statistics*. 28 (2), 461–482.

